

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 07-271738

(43)Date of publication of application : 20.10.1995

(51)Int.Cl.

G06F 15/16

G06F 9/46

(21)Application number : 07-041126

(71)Applicant : NEC CORP

(22)Date of filing : 28.02.1995

(72)Inventor : SURETSUSHIYU JIYAGANASAN
JIEEMUSU EFU FUJIRUBIN

(30)Priority

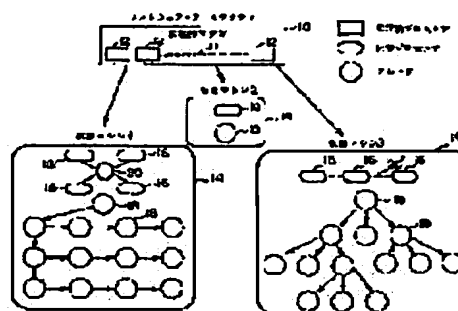
Priority number : 94 221026 Priority date : 31.03.1994 Priority country : US

(54) CONTROL SYSTEM FOR HIGH-LEVEL PARALLEL COMPUTER SYSTEM USING SOFTWARE ARCHITECTURE

(57)Abstract:

PURPOSE: To provide the control system which uses the software architecture, equipped with several layers of abstract bodies, controlling the high-level parallel computer system.

CONSTITUTION: An abstract physical machine 10 (1st layer) includes a group of abstract physical processors and is considered as a microkernel. A 2nd layer includes a virtual machine 2 and virtual processors 16. The virtual machine is equipped with a virtual address space and a group of virtual processors connected by virtual topology. The virtual machine is mapped to the abstract physical machines and each virtual processor is mapped to the abstract physical processors. A 3rd layer defines threads 18. The threads are processes with light weight which run on the virtual processors. The abstract physical machine, abstract physical processors, virtual machines, virtual processors, thread groups, and threads are preferably all objects of a first class.



LEGAL STATUS

[Date of request for examination] 01.03.1995

[Date of sending the examiner's decision of rejection] 27.05.1998

[Kind of final disposal of application other than the
examiner's decision of rejection or application
converted registration]

[Date of final disposal for application]

[Patent number] 2848262

[Date of registration] 06.11.1998

[Number of appeal against examiner's decision of
rejection] 10-10006

[Date of requesting appeal against examiner's] 25.06.1998

decision of rejection]
[Date of extinction of right]

(19) 日本国特許庁 (J P)

(12) 公 開 特 許 公 報 (A)

(11) 特許出願公開番号

特開平7-271738

(43) 公開日 平成7年(1995)10月20日

(51) Int.Cl. ⁴	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 15/16	4 3 0 C			
9/46	3 5 0	7737-5B		

審査請求 有 請求項の数20 O L (全 30 頁)

(21) 出願番号 特願平7-41126

(22) 出願日 平成7年(1995)2月28日

(31) 優先権主張番号 08/221026

(32) 優先日 1994年3月31日

(33) 優先権主張国 米国 (U S)

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72) 発明者 スレッシュ ジャガナサン

アメリカ合衆国, 95134 カリフォルニア,

サン ジョーズ, リオ ロブルス 110

エヌ イー シー アメリカ, インコーポ
レイテッド内

(72) 発明者 ジェームス エフ. フィルビン

アメリカ合衆国, 95134 カリフォルニア,

サン ジョーズ, リオ ロブルス 110

エヌ イー シー アメリカ, インコーポ
レイテッド内

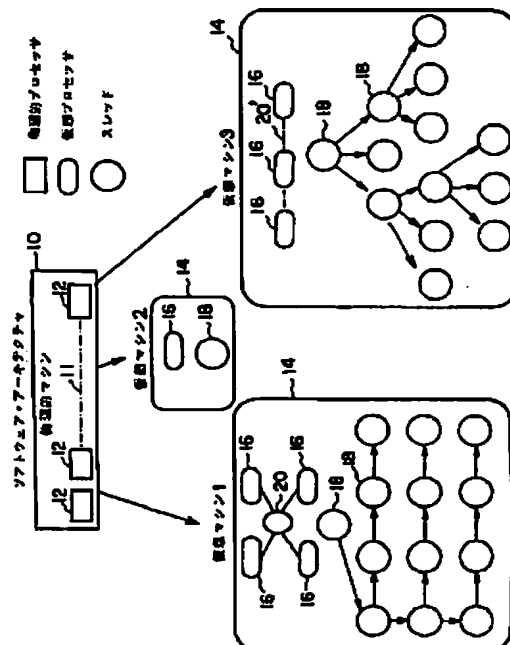
(74) 代理人 弁理士 後藤 洋介 (外2名)

(54) 【発明の名称】 ソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式

(57) 【要約】

【目的】 いくつかの抽象体の層を備えた、高度並列コンピュータ・システムを制御するソフトウェア・アーキテクチャを用いた制御方式の提供。

【構成】 抽象物理的マシン10 (第1層) は抽象物理的プロセッサの組を含んでおり、マイクロカーネルと考えることができる。第2層は仮想マシン2と仮想プロセッサ16とを含んでいる。仮想マシンは仮想アドレス空間と、仮想トポロジーで接続された仮想プロセッサの組とを備えている。仮想マシンは抽象物理的マシンにマッピングされ、各仮想プロセッサは抽象物理的プロセッサにマッピングされている。第3層は、スレッド18を定義している。スレッドは、仮想プロセッサ上でランするライトウエイトのプロセスである。望ましくは、抽象物理的マシン、抽象物理的プロセッサ、仮想マシン、仮想プロセッサ、スレッド・グループ、ならびにスレッドはすべてファーストクラスのオブジェクトである。



【特許請求の範囲】

【請求項1】 高度並列コンピュータ・システムを制御するためのソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式において、一つのマイクロカーネルを形成する複数の抽象物理的プロセッサを備えた複数の抽象物理的マシンと；前記複数の抽象物理的プロセッサに付随し、複数の仮想プロセッサを備えた複数の仮想マシンと；前記複数の仮想プロセッサ上でランする複数のスレッドを備えた複数のスレッド・グループとを備え、前記複数の仮想プロセッサおよび前記複数のスレッドはファーストクラスのオブジェクトであることを特徴とするソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項2】 前記複数の仮想プロセッサは仮想トポロジにおいて接続されていることを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項3】 前記マイクロカーネルのポリシーを管理するマイクロカーネル・ポリシー・マネージャはユーザがカスタマイズできることを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項4】 前記複数の仮想プロセッサは、前記複数のスレッドのポリシーを管理する複数のスレッド・ポリシー・マネージャのうち、ユーザが、どのスレッド・ポリシー・マネージャをカスタマイズできるかを含むことを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項5】 前記複数のスレッド、前記複数の仮想プロセッサ、ならびに前記複数の抽象物理的プロセッサは、機能的に連携し、仮想トポロジを構築することを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項6】 前記仮想トポロジはユーザがカスタマイズできることを特徴とする請求項5記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項7】 前記複数のスレッドは、それらのそれぞれの実行コンテキストから分離でき、実行コンテキストの遅延された割り当てを許すことを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項8】 複数の多様な形態のポートをさらに備えたことを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項9】 前記複数の多様な形態のポートはそれぞ

れファーストクラスのオブジェクトであることを特徴とする請求項8記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項10】 前記複数のスレッドは、一般データと複合データとを含むメッセージを送ることを特徴とする請求項8記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項11】 前記複数のスレッドは、それらのそれぞれのローカル・スタックおよびヒープを、他の複数のスレッドとは独立に、ガーベッジ・コレクトすることを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項12】 前記複数のスレッド・グループはそれらのそれぞれの共有ヒープを、無関係の複数のスレッド・グループとは独立に、集めることを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項13】 前記複数の仮想プロセッサは前記複数の抽象物理的プロセッサ上に多重化されていることを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項14】 前記複数の仮想プロセッサ、前記複数の仮想マシン、ならびに前記複数のスレッドは、持続性メモリ内に存在することを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項15】 前記複数の抽象物理的プロセッサはファーストクラスのオブジェクトであることを特徴とする請求項1記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項16】 前記複数の仮想マシンはファーストクラスのオブジェクトであることを特徴とする請求項15記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項17】 前記複数の抽象物理的マシンおよび前記複数のスレッド・グループはファーストクラスのオブジェクトであることを特徴とする請求項16記載のソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式。

【請求項18】 各々が仮想プロセッサ・コントローラと仮想プロセッサ・ポリシー・マネージャとを有し、物理的トポロジにおいて接続された複数の抽象物理的プロセッサと；各々が、仮想アドレス空間と複数の仮想プロセッサとを有する複数の仮想マシンと；を備えたコンピュータ・システムであって、

前記複数の仮想マシンの各々の前記複数の仮想プロセッサは、前記仮想プロセッサ・コントローラ及び前記仮想プロセッサ・ポリシー・マネージャに応答して実行し、かつ、スレッド・コントローラとスレッド・ポリシー・

マネージャとを有し、前記複数の仮想プロセッサは仮想トポロジーにおいて接続され、各仮想プロセッサはそれぞれの抽象物理的プロセッサにマッピングされており、前記コンピュータ・システムは、前記スレッド・コントローラと前記スレッド・ポリシー・マネージャとにตอบสนองする前記複数の仮想プロセッサ上でランする複数のスレッドを、更に、備えていることを特徴とするコンピュータ・システム。

【請求項19】 前記複数の仮想プロセッサは前記複数の抽象物理的プロセッサ上で多重化されていることを特徴とする請求項18記載のコンピュータ・システム。

【請求項20】 更に、持続性メモリを備え、この持続性メモリには、前記複数のスレッド、前記複数の仮想プロセッサ、ならびに前記複数の仮想マシンを含むオブジェクトが存在することを特徴とする請求項18記載のコンピュータ・システム。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は、高度に並列化したコンピュータ・システムを制御するためのコンピュータ・ソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式に関し、特に、現代のプログラミング言語に対して、非常に効率的なサブストレートとして役立つよう設計したコンピュータ・ソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式に関するものである。

【0002】このコンピュータ・ソフトウェア・アーキテクチャを用いた制御方式は、制御の問題をポリシーの問題から分離したオペレーティング・システムにもとづいている。この分離はシステムの2つの異なる抽象レベルで行っている。すなわち抽象物理的プロセッサと仮想プロセッサとにおいてである。これら抽象体のそれぞれは2つの構成要素に分れている。1つは、抽象体の制御部分を実現する“コントローラ”であり、もう1つはコントローラに対してポリシーを決定する“ポリシー・マネージャ”である。制御をポリシーから分離することによって、機能的に同一のシステムに対する異なる振舞の定義を、抽象体のポリシー・マネージャ部分を変更するのみで行うことが可能となる。

【0003】具体的には、このソフトウェア・アーキテクチャを用いた制御方式は、制御のライトウエイト・スレッドと仮想プロセッサとをファーストクラスのオブジェクトとしてサポートする。並行（コンカレンシー）マネージメントは、ファーストクラスのプロシージャおよびファーストクラスのコンティニュエーションによって実現する。それによって、ユーザが基本的なランタイム・システムに関する知識を持っていなくても、アプリケーションのランタイムの振舞を最適化することが可能となる。

【0004】さらに具体的には、本発明は、非同期並行

構造の構築と、基本的制御メカニズムとしてコンティニュエーションを用いてスレッド・コントローラの実現と、大規模並行計算の組織化と、並列計算のための強固なプログラミング環境との設計に関するものである。

【0005】

【従来の技術】並列計算に対する興味が高まり、その結果、並行性を表現するために高レベルのプログラムとデータの構造を明確に定義する並列プログラミング言語が多数、生み出された。非数値的アプリケーション領域をターゲットにする並列言語は典型的に、動的なライトウエイト・プロセスの生成、高レベル同期基本命令、分散データ構造、ならびにスペキュラティブな並行性を実現する並行構造をサポートする（効率とは異なっている）。これらの並列言語は事実上すべて、2つの部分言語から成ると考えられる。2つの部分言語とは、すなわち、プロセスの集まりのアクティビティを管理し同期化する調整言語と、与えられたプロセスに限定されるデータ・オブジェクトを扱う計算言語とである。

【0006】伝統的に、オペレーティング・システムにはいくつかのクラスがある。例えば、リアルタイム、インタラクティブ（会話型）、バッチなどである。これら3つのクラスのオペレーティング・システムはユーザに対して異なるインターフェースを提供するので、プログラムをあるクラスのオペレーティング・システム（OS）から他のクラスのオペレーティング・システムに移動するのは困難であった。さらに、各クラスのオペレーティング・システムが決定するスケジューリングは異なっているので、1つのオペレーティング・システムのためのプログラム（例えばリアルタイム・アプリケーション）を他のオペレーティング・システム（例えば会話型開発システム）でデバッグすることは難しく、またアプリケーションがターゲット・システム上で正確かつ効率的にランすることに関して自信を持つことは困難である。

【0007】そして、これらのクラスのシステムではそれぞれ異なるスケジューリング方式を用いているので状況はさらに複雑である。例えば、ある種のリアルタイム・システムでは、複数のプロセスに対してスケジューリングの順序は固定しているのに対して、別のシステムでは優先規律を用いたり、ランニング・クオンタムを用い、さらに他のシステムでは、それらを組み合わせている。会話型オペレーティング・システムあるいはバッチ・オペレーティング・システムはスケジューリングに関してかなりの数の選択肢を有している。

【0008】制御をポリシーから分離することによって、種々のクラスのオペレーティング・システムに対して容易にカスタマイズできるオペレーティング・システムを構築できる。本発明では、ポリシー・マネージャを実現するモジュールは典型的にはシステムのサイズに比べて非常に小さい。一般にコードのライン数は100未

満である。従って、ポリシーの振舞が異なるシステムを新たに構築する場合、通常はコードの小部分を書くのみでよい。また、ポリシー・マネージャは良く定義されたインターフェースを提供するので、ポリシーの振舞を変更した場合、システム全体を試験する必要はなく、新しいポリシー・マネージャだけを試験すればよい。

【0009】Hydra (参考文献: "HYDRA/C. mmpi: An Experimental Computer System", William Wulf, Roy Lexia, 及び Samuel Harblson 著, McGraw-Hill, 1991) は、制御とポリシーとの分離を意図して設計された最初のオペレーティング・システムである。しかし、Hydra はポリシーのカスタマイズをカーネルのレベルでしか認めていない。本発明ではさらに進めて、ポリシーの決定を、それらが特定のプログラムに関連するものである場合、カスタマイズできるようにする。従って、エディタやウィンドウ・マネージャなどの会話型プログラムは、流体力学のシミュレーションや有限要素法の計算など、計算を主体とするプログラムとは非常に異なったポリシーを持つことができる。また、Hydra における制御とポリシーとの分離はコストのかかるものとなっている。それは、カーネルとポリシー・マネージャとの間に複数のコンテキスト・スイッチを必要とするからである。本発明では、ポリシー・マネージャは一般に適当なアドレス空間に直接リンクしており、コンテキストの切り換えは不要である。従って本発明のポリシー・マネージャは少なくとも従来のオペレーティング・システムにおけるポリシー・マネージメント(カスタマイズできない)と同程度に効率的であり、そして通常は従来以上に効率が良い。

【0010】高レベルの並列言語を実現する1つの方法は、専用の(ユーザ・レベル)仮想マシンを構築することである。仮想マシンは基本的に、調整部分言語に見られる高レベルの並行プリミティブを実現するサブストレートとして機能する。調整言語Lが並行プリミティブPをサポートする場合、Lの仮想マシン(L_r)の役割は、Pの実現に関連したことをすべて扱うことである。そのためにはマシンが、プロセスのスケジューリング、記憶、管理、同期化、その他を管理することがしばしば必要となる。しかし、L_rはPを効率良く実現するようにのみ調整されているので、非常に異なった並行プリミティブを実現することは多くの場合適当でない。従って、並行プリミティブP'によってL_rの方角を構築するためには通常、仮想マシンを新たに構築するか、あるいはPを用いてP'の意味規制を表現する必要がある。これら2つのアプローチには明らかに欠点がある。すなわち、第1のアプローチでは複雑な仮想マシンを新たに構築するため、コスト高である。一方、第2のアプローチは、Pの意味規制が高レベルであり、またL_rの機能が

限定されているので、不十分である。

【0011】言語の実現において、並行性を実現するために専用の仮想マシンを構築する代りに、低レベルのオペレーティング・システムの機能を用いることができる。プロセスの生成およびスケジューリングは、OSが管理する、制御のスレッド(ヘビーウエイトあるいはライトウエイト)によって実現する。そして同期化は、低レベルの、OSが管理する構造体を用いて扱う。このようにして実現したものは一般に、専用の実行時システムの周辺に構築したシステムより、ポータブルであり、また拡張性が高い。ただし、カーネル(低レベル)はすべて、アプリケーションとオペレーティング・システムとの間の保護境界を横断する必要があるため、効率は犠牲になる。さらに、汎用のOS機能は通常、対象の並行オペレータの意味規制に対して不感であるため、それらはコンパイル時間あるいは実行時間の点で最適化をほとんど、あるいはまったく行わない。

【0012】

【発明が解決しようとする課題】高度並列マルチプロセッサ/マルチコンピュータ・システムを制御するための、現代のプログラミング言語に対する非常に効率の良いサブストレートとして役立つコンピュータのオペレーティング・システム・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。更に、本発明によれば、カスタマイズ可能な仮想マシンにもとづく非同期の計算のためのソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0013】また、本発明によれば、仮想プロセッサ上でファーストクラスのオブジェクトとしてライトウエイト・スレッドをサポートするソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0014】更に、本発明によれば、カスタマイズ可能なポリシー・マネージャを、特にユーザ・レベルに含むソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0015】また、本発明によれば、カスタマイズ可能な仮想トポロジーを含むソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0016】また、本発明によれば、スレッド吸収、遅延TCB割り当て、ならびに記憶装置共有の場所としてのスレッド・グループを含むソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0017】更に、本発明によれば、多様な形態のポートを含むソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0018】また、本発明によれば、上述のようなソフ

トウェア・アーキテクチャを用いて制御されるコンピュータ・システムが得られる。

【0019】

【課題を解決するための手段】本発明によれば、高度並列コンピュータ・システムを制御するためのソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式において、一つのマイクロカーネルを形成する複数の抽象物理的プロセッサを備えた複数の抽象物理的マシンと；前記複数の抽象物理的プロセッサに付随し、複数の仮想プロセッサを備えた複数の仮想マシンと；前記複数の仮想プロセッサ上でランする複数のスレッドを備えた複数のスレッド・グループとを備え、前記複数の仮想プロセッサおよび前記複数のスレッドはファーストクラスのオブジェクトであることを特徴とするソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0020】更に本発明によれば、各々が仮想プロセッサ・コントローラと仮想プロセッサ・ポリシー・マネージャとを有し、物理的トポロジにおいて接続された複数の抽象物理的プロセッサと；各々が、仮想アドレス空間と複数の仮想プロセッサとを有する複数の仮想マシンと；を備えたコンピュータ・システムであって、前記複数の仮想マシンの各々の前記複数の仮想プロセッサは、前記仮想プロセッサ・コントローラ及び前記仮想プロセッサ・ポリシー・マネージャに依存して実行し、かつ、スレッド・コントローラとスレッド・ポリシー・マネージャとを有し、前記複数の仮想プロセッサは仮想トポロジにおいて接続され、各仮想プロセッサはそれぞれの抽象物理的プロセッサにマッピングされており、前記コンピュータ・システムは、前記スレッド・コントローラと前記スレッド・ポリシー・マネージャとに依存する前記複数の仮想プロセッサ上でランする複数のスレッドを、更に、備えていることを特徴とするコンピュータ・システムが得られる。

【0021】本発明は、高レベル・プログラミング言語のコンテキストにおいて広い範囲の並行構造体を表現することを可能とする調整サブストレートの実現に関するものである。本発明は汎用調整モデルを定義し、そのモデル上で、多数の特殊調整言語を効率良く実現できるようにする。本発明の実施においては、ソフトウェアのスキーム (Scheme) (参考文献: "The Revised Report on the Algorithmic Language Scheme" ACM Sigplan Notices, 21 (12), 1986, Jonathan Rees and William Clinger) を計算の基礎として用いた。スキームはより高次の、辞書的に見たときの、Lisp の方言である。スキームは望ましい言語ではあるが、当業者にとって明らかなように、上記調整サブストレートの設計は、いかなる現代の (高レベルの) プログ

ラミング言語にも取り入れることができよう。

【0022】本発明のオペレーティング・システムは基本的に、共有メモリあるいは分離メモリを用いた、MIMD (マルチ・インストラクション・マルチ・データ) 並列コンピュータ上でランするように設計され、またワークステーションのネットワークから成る分散マシン上でランするように設計されている。本発明のソフトウェア・アーキテクチャでは、分離メモリあるいは分散メモリを用いたマシン上で実行する場合には、共有仮想メモリ・モデルを用いる。その実現においては、異なる、並列のパラダイムに対応する多数の異なるアルゴリズムを用いた。上記並列には結果並列、マスター/スレーブ並列、ならびに論理的並列が含まれる。いくつかの異なる並列プログラミング・モデルを、フューチャー・ファーストクラスのタプル (組) 空間、ならびにエンジンを備えたオペレーティング・システム上で実現した。

【0023】本発明の望ましい実施例のフィーチャーである、オペレーティング・システム (OS) を構成するスキームの方言 (スティング (Sting) と呼ぶ) は、非同期、ライトウエイトの並行性を表現するための調整言語 (専用仮想マシンによって実現) を含み、それは2つのアプローチの最良点を組み合わせている。他の並列スキームのシステムおよび同種の高レベル言語の並列方言と異なり、スティングにおける基本的な並行オブジェクト (スレッド、仮想プロセッサ、ならびに物理的プロセッサ) は、ストリームライン化したデータ構造であり、複雑な同期化を必要としない。並行性の管理をOSによるサービスに依存する並列システムと違い、スティングはスキームのオブジェクトおよびプロシージャによってすべての並行管理の問題を実現し、その結果、ユーザは、背後のOSのサービスに関する知識を持つことなく、アプリケーションのランタイムの振舞いを最適化することが可能となる。スティングは、さまざまな形態の非同同期の並列性を生成し、管理するための基本的な特徴を、概念的に単一化したフレームワークで、かつ非常に一般的なフレームワークによってサポートする。結果として、高レベル言語の種々の並列方言をその上に構築できる効率的なサブストレートを構築できることが分った。スティングは単に、スタンドアロンの、短寿命のプログラムを実現する媒介手段とすることを意図したものではなく、並列計算のための豊かなプログラミング環境を構築するためのフレームワークを提供することを期待したものである。従って、このシステムは、スレッド・プリエンブション、スレッドごとの非同同期のガーベッジ・コレクション、スレッド境界を越えた例外の扱い、ならびにアプリケーションに依存するスケジューリング・ポリシーをサポートする。さらに、このシステムは、持続性の長寿命なオブジェクト、マルチ・アドレス空間、その他、最新のプログラミング環境に共通する特徴を扱うために、必要な機能を有している。

【0024】スティングでは、仮想プロセッサは抽象物理的プロセッサ上で多重化され、スレッドは仮想プロセッサ上で多重化される。この多重化に関連するポリシーの決定はすべて、ポリシー・マネージャによって行われる。物理的プロセッサ上の仮想プロセッサの多重化に関連する決定は、仮想プロセッサ・ポリシー・マネージャ（VPPM）によって行う、仮想プロセッサ上のスレッドの多重化に関する決定はスレッド・ポリシー・マネージャ（TPM）によって行われる。

【0025】ポリシー・マネージャは3つのタイプの決定を行う。すなわち、オブジェクトが生成あるいは再開されたとき、プロセッサ（物理的あるいは仮想）に新しいオブジェクト（VPあるいはスレッド）をいかにマッピングするか、特定のプロセッサにマッピングされた複数のオブジェクトをランさせる順序、ならびにオブジェクトをあるプロセッサから他のプロセッサに、いつ再マッピングあるいは移動するか3つである。

【0026】スティングは、スキーム、スモールトーク（Small Talk）、ML、モジュール3（Module 3）、あるいはハスケル（Haskell）などの現代のプログラミング言語をサポートするように設計されたオペレーティング・システムである。スティングは、低レベルの直交構築体の基礎を与え、それによって言語の設計者あるいは使用者が、上記言語が必要とする種々の構築体を簡単かつ効率的に構築することを可能とする。

【0027】現代のプログラミング言語は、従来のコボル、フォートラン、C、あるいはバスケルなどのプログラミング言語に比べ、より多くを要求する。スティングは現代のプログラミング言語をサポートするように設計されているが、従来のプログラミング言語も同様に効率良くサポートする。現代のプログラミング言語が従来の言語と異なる点を以下にリストアップする。

【0028】・並列性：汎用のマルチ・プロセッサはますます利用し易くなってきており、その結果、並行プログラミングのための効率的で、かつ表現力に優れたプラットフォームの構築に対して興味が高まっている。高レベルのプログラミング言語に並行性を組み入れるための努力は大部分が、特殊目的の基本命令を言語に付加するという点に払われている。

【0029】・マルチ同期化モデル：並列プログラミングあるいは非同期プログラミングにおいて、多くの同期化プロトコルが用いられている。現代のオペレーティング環境は、できる限りさまざまなプロトコルをサポートする基本命令を提供するものでなければならない。

【0030】・レイジー（遅延）評価およびイーガー評価：現代の多くの言語はレイジー評価あるいはイーガー評価のいずれか、または両方をサポートしている。オペレーティング・システムにとって、レイジーからイーガーまでの完全な評価ストラテジーを用意することは重要

である。

【0031】・自動記憶管理：これは現代の多数の言語の基本的な特徴となっている。それは、自動記憶管理によってプログラムを一層、表現力に優れたものにでき、同時にプログラムのエラーを低減し、かつプログラムの複雑さを緩和できるからである。

【0032】・トポロジー・マッピング：多くのプログラミング言語ではまだサポートされていないが、プログラムにおける通信オーバーヘッドを低減するように、処理のプロセッサへのマッピングを制御する能力は、マルチ・プロセッサ・コンピュータ・システムのサイズが大きくなり続け、かつトポロジーがより複雑になる以上、より重要なものとなろう。

【0033】スティングはこれら種々の要素を効率良くサポートする。スティングは、現在利用できるものより一層、一般的でかつより効率的なアーキテクチャ・フレームワークにおいてこれを行う。スティングはまた、高い表現力および制御能力と、非並列レベルのカスタマイズ能を、プログラムに提供する。

【0034】スティングは、その設計における4つの特徴によって、他の並列言語から最もよく区別できる。

【0035】1. 並行抽象体：並行性はスティングでは制御のライトウエイト・スレッドによって表現される。スレッドは非疎密な、ファーストクラスのデータ構造である。

【0036】2. プロセッサ抽象体およびポリシー抽象体：スレッドは、スケジューリングおよび負荷平衡・プロトコルの抽象体を表す仮想プロセッサ（VP）上で実行する。仮想プロセッサの数は、実際に利用できる物理的プロセッサの数より多くてもかまわない。スレッドのように、仮想プロセッサはファーストクラスのオブジェクトである。1つのVPは1つのスレッド・ポリシー・マネージャを備え、このポリシー・マネージャはそれが実行するスレッドのためのスケジューリングと移行方式を決定する。異なるVPは、実際には、性能の低下無しに、異なるポリシー・マネージャを備えることができる。仮想プロセッサは、実際の物理的計算装置である物理的プロセッサ上で実行する。

【0037】仮想プロセッサの集まりとアドレス空間とは組合わさって、1つの仮想マシンを形成する。複数の仮想マシンが単一の物理的マシン上で実行できる。物理的マシンは1組の物理的プロセッサから成る。仮想マシンおよび物理的マシンもまた指示可能な、スキームのオブジェクトであり、このオブジェクトとして操作可能である。

【0038】3. 記憶モデル：1つのスレッドはデータを、そのスレッドが排他的に管理するスタックおよびヒープに割り当てる。従って、複数のスレッドは、互いに独立にそれらのプライベート・ステートのガーベッジ・コレクションを行う。あるスレッドがプライベートのガ

ページ・コレクションを開始する場合、グローバルな同期化は不要である。データはスレッドを横断して参照できる。スレッド境界を越えてオブジェクトのガーベージ・コレクションを行うとき、領域間の参照情報が用いられる。記憶は世代スキッピング・コレクタによって管理される。1つのスレッドによって割り当てられた長寿命データあるいは持続データは、同じ仮想マシンにおける他のスレッドもアクセスできる。

【0039】本発明の設計は記憶のローカルリティということに配慮している。例えば、スレッドをランさせるための記憶装置はVPにキャッシュされ、そして1つのスレッドが終了したとき、すぐに再利用できるようリサイクルされる。さらに、複数のスレッドは、データの依存性が保証されるときは常に、同じ実行コンテキストを共有することができる。

【0040】4. プログラム・モデル：スティングは、スレッド間で横断的に扱われるべき例外を許容し、ノン・ブロッキングI/Oをサポートし、仮想プロセッサのスケジューリングのカスタマイズを、仮想プロセッサ上のスレッドのスケジューリングのカスタマイズ可能であるのと同様に、可能とし、そしてマルチ・アドレス空間および共有持続オブジェクトを実現する内部構造を与える。スティングはまた、ファーストクラスの多様な形態のポートを用いたメッセージの効率の良い受け渡しをサポートする。ポートは、分離メモリ・プラットフォーム上の共有メモリの実現において、オーバーヘッドを緩和するのに役立つ。

【0041】本発明の高度並列コンピュータ・システムを制御するソフトウェア・アーキテクチャでは、オペレーティング・システム（スティング）、基本言語、ならびにコンパイラを1つの抽象的マシンに統合する。スタート点はスキームなどの高レベルプログラミング言語である。このプログラミング言語は、スレッド、仮想プロセッサ、ならびにポリシー・マネージャを含む効率的な抽象体によって拡大されている。この優れたオペレーティング・システムは、データのローカルリティにプレミアムを付けるという現在のアーキテクチャのトレンドを有効に利用したメカニズムを含んでいる。

【0042】その結果、並列計算のための効率の良い調整構造体を構築するメカニズムが得られた。ライトウェイトのスレッドを用いることにより、進歩的なプログラミング環境の基礎が得られる。データのローカルリティをサポートすることによって、効率的な非同期システムが得られる。

【0043】このシステムの性能にとって中心的なことは仮想トポロジーの概念である。仮想トポロジーは、仮想プロセッサの集まりにおける関係を定める。ツリー、グラフ、ハイパーキューブ、ならびにメッシュとして構成されたプロセッサ・トポロジーはよく知られたその例である。仮想プロセッサは、スレッドが実行するスケジ

ューリング、マイグレーション、ならびに負荷平衡のポリシーを定義する抽象体である。この仮想トポロジーは、複雑なスレッドとプロセッサのマッピング（物理的相互接続の低レベルの詳細を抽象する）を定める、単純で表現力に優れた高レベルのフレームワークを与えるよう意図されている。

【0044】計算によって生成されたスレッドは、仮想トポロジー内のプロセッサに対して、そのトポロジーに関連したマッピング機能によってマッピングされる。ユーザはこれらのマッピング機能を定義することができる。仮想トポロジーを用いて特定のマルチプロセッサ・プラットフォーム上でシステムが実現されている場合、仮想トポロジー内の仮想プロセッサをプラットフォーム内の物理的プロセッサにマッピングするプロシージャを定義することが可能である。

【0045】コードそれ自身は、それが物理的プロセッサあるいは物理的プロセッサの相互接続に対する参照を含んでいない限り、マシンとは独立している。スレッド・マッピングとローカルリティに関するすべてのことは、プログラムが用いる、仮想トポロジーの仕様と、プログラム実行時のトポロジー内のノードの通過の仕方において抽象される。

【0046】仮想トポロジーとプロセッサ・マッピングの利益は、効率性だけでなく、移植性という点にもあり、それによって並列アルゴリズムの実現を個別の物理的トポロジーごとに特殊化する必要がなくなる。スレッドをプロセッサに関連づけるマッピング・アルゴリズムは、仮想トポロジーの一部として細かく指定されるので、プログラマは、スレッドがどのように仮想プロセッサに対してマッピングされるべきかを正確に管理できる。ある計算において通信が必要となることが分かっている場合、これらのスレッドを特定の仮想プロセッサに明確に割り当てられるという能力によって、暗黙的なマッピング・ストラテジーの場合より優れた負荷平衡を行える。並列アルゴリズムによって定義される制御とデータフローのグラフの構造は、種々の形で用いることができる。スレッドの集まりが共通のデータを共有している場合には、これらのスレッドが実行する仮想プロセッサを同一の物理的プロセッサにマッピングするトポロジーを構築することが可能である。仮想プロセッサは物理的プロセッサ上で、スレッドが仮想プロセッサ上で多重化されるのと同じようにして多重化される。あるスレッドの集まりが重要な相互の通信を必要とする場合には、それらのスレッドを、仮想トポロジーにおいて互いに接近したプロセッサにマッピングするトポロジーを構築することができる。スレッド T_1 が、他のスレッド T_2 が発生する値に対してデータ依存性を有している場合、 T_1 と T_2 とは同一の仮想プロセッサにマッピングすることが合理的である。プロセッサがほとんどビジー状態となるグラニュラリティの細かいプログラムでは、同一また

13

は近いプロセッサ上のデータ依存スレッドに対してスケジューリングを行える能力によって、スレッドのグラニュラリティを改善する機会が与えられる。最後に、適応ツリー・アルゴリズムなど、ある種のアルゴリズムは計算の進行につれて展開するというプロセス構造を有している。これらのアルゴリズムは、仮想プロセッサの動的生成が可能なトポロジー上において最も良く実行される。

【0047】このソフトウェア・アーキテクチャの他の優れた面として、効率的な汎用のマルチ・スレッドのオペレーティング・システムおよびプログラム環境の実現における、コンティニュエーションおよびファーストクラスのプロシージャの役割がある。コンティニュエーションは、状態遷移の操作、例外の扱い、ならびに重要な記憶の最適化を実現するために用いられる。コンティニュエーションは、プログラム・ポイントの抽象体である。コンティニュエーションは、1つの引数を有するプロシージャによって表され、このプロシージャは、引数が示すプログラム・ポイントから実行すべき残りの計算を定義している。

【0048】スティングの仮想アドレス空間は1組の領域によって構成されている。領域は、一時的にあるいは空間的に強いローカリティを示すデータを組織化するために用いられる。スティングはさまざまな領域をサポートする。すなわち、スレッド制御ブロック、スタック、スレッド・プライベート・ヒープ、スレッド共有ヒープなどである。データは、それらの意図された仕様および寿命にもとづいて領域に割り当てられ、従って異なる領域は、それらに関連した異なるガーベッジ・コレクタを備えることになる。

【0049】例外と割り込みは常に、スレッド・レベルのコンテキスト・スイッチの場合のように、あるスレッドの実行コンテキストにおいて扱われる。例外ハンドラーは通常のスキームのプロシージャによって実現され、そして例外のディスパッチは基本的にコンティニュエーションの操作を含んでいる。

【0050】スティングが、制御のライトウエイト・スレッドの生成および管理が可能なプログラミング・システムである限り、いくつかの特性を、他の高レベル言語のために開発されたスレッド・パッケージ・システムと共有している。これらのシステムもスレッドを明らかなデータタイプと見ており、また、さまざまな程度にプリエンプションをサポートし、そしてある限定されたケースでは、プログラマが特別のスケジュール管理を指定することを可能としている。これらのシステムでは、スレッドの抽象体が調整部分言語を定めている。

【0051】しかし、スティングはいくつかの重要な点でこれらのシステムと異なっている。第1に、スティングが使用するスケジューリングとマイグレーションのプロトコルは完全にカスタマイズできる。異なるアプリケ

14

ーションは、スレッド・マネージャあるいは仮想プロセッサの抽象体を変更することなく、異なるスケジューラをランさせることができる。このようなカスタマイズは仮想マシン自身の組織化に適用することができる。第2に、スティングによるデータのローカリティのサポート、記憶の最適化、ならびにスレッドの吸収によるプロセスの抑圧は他のシステムでは行えない。さらに、スレッドのオペレーションはすべてスレッドの仮想マシン内で直接実現される。スレッドのオペレーションの実行のために実施すべき、低レベルのカーネルに対するコンテキスト・スイッチは無い。スティングは、長寿命のアプリケーション、持続性のオブジェクト、ならびにマルチ・アドレス空間をサポートすることを意図した抽象的マシンにおいて構築される。スレッド・パッケージは、それらが（定義によって）完全なプログラム環境を定めていないので、これらの機能はまったく提供しない。

【0052】スティングはシステム・プログラミング言語として設計されているので、低レベルの平行抽象体を提供する。アプリケーション・ライブラリは直接スレッド・オブジェクトを生成でき、そしてそれら自身のスケジューリングおよびスレッド・マイグレーション・ストラテジーを定めることができる。高レベルの平行構築体はスレッドを用いて実現できるが、しかし効率が保証されるなら、システムはユーザがスレッドのオペレーションを上述のように直接利用することを禁止するものではない。具体的には、同一のアプリケーションは、同一の実行時の環境において、異なる意味規制と異なる効率で、平行抽象体を定めることができる。

【0053】ある点でスティングは、他の進歩的マルチ・スレッド・オペレーティング・システムに似ている。例えば、スティングは、コール・バック、ユーザが管理するオーバー・インタラプト、ならびにユーザ・レベルの操作としてのローカル・アドレス空間の管理に伴うノンブロッキングI/Oコールをサポートしている。スティングはユーザ・レベルの事柄とカーネル・レベルの事柄とを分けている。物理的プロセッサは（特権を与えられた）システムのオペレーション、および複数の仮想マシンに跨るオペレーションを扱う。仮想プロセッサはユーザ・レベルのスレッドおよびローカル・アドレス空間の機能をすべて実現する。しかし、スティングはスキームの拡張方法であるため、典型的なオペレーティング・システム環境では提供されない高レベルのプログラミング言語の機能性および表現性を提供する。

【0054】スティングは、非同期プログラミング基本命令を構築し、そして新しい並列プログラミングのパラダイムを実験するためのプラットフォームである。さらに、その設計では、異なる平行性の手法を競走的に評価することが可能である。スキームは、意味規制が良く定義され、全体的に簡素であり、そして効率的であるため、このような実験を行うための特に豊かな環境を提供

する。しかし、スティングの設計はそれ自身言語に依存しない。従って、いかなる高レベルプログラミング言語にも極めて容易に組み込むことができよう。

【0055】スティングは単に、興味深いと思われる各平行パラダイムおよび各平行プリミティブに対してフックを与えるものではない。そうではなく、広範囲の並列プログラミング構造体に共通の基本構造および機能に焦点を当てている。従って、ブロッキングの実現は論理的な計算をサポートするために容易に用いられる。スレッドの実行を抑止するために用いられるスレッド吸収の最適化は、フューチャーとタプル空間の同期化を実現するのに非常に適しており、そして最後に、カスタマイズ可能なポリシー・マネージャは、他のさまざまなパラダイムに対して公正で効率的なスケジューラを構築することを可能とする。

【0056】

【実施例】次に本発明の実施例について図面を参照して説明する。

【0057】図1に本発明の一実施例による高度並列コンピュータ・システムを制御するためのソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式のブロック図を示す。

【0058】抽象物理的マシン(PM)10は、物理的トポロジー(PT)11で互いに接続された抽象物理的プロセッサ(PP)12により構成されている。この抽象物理的マシンは1組の仮想マシン(VM)14を実行させるために用いられる。それに対して、各仮想マシンは、仮想トポロジー(VT)20、20'で接続された1つ以上の仮想プロセッサ(VP)16を備えている。スレッド(T)18は、同じ仮想マシン内の1つ以上の仮想プロセッサ上で実行する。さらに、特定のスレッドは、同じ仮想マシン14内の異なる仮想プロセッサ間で移行(マイグレート)できる。スレッド・ポリシー・マネージャ(TPM)19(図2、図3に示す)はスレッドのスケジューリングおよびスレッドの負荷平衡・ポリシーを制御する。異なる要素間の関係および各要素の詳細を以下に説明する。

【0059】ソフトウェア・アーキテクチャ(オペレーティング・システム・アーキテクチャという場合もある)は、いくつかの抽象体の層の配列と考えることができる(図2)。第1の層は抽象物理的マシン10を含み、このマシンは抽象物理的プロセッサ12の組を含んでいる。この層は、現状のオペレーティング・システムにおいてマイクロ・カーネルと呼ばれているものに対応している。次の層は仮想マシン14および仮想プロセッサ16を含んでいる。仮想マシンは、仮想アドレス空間と、仮想トポロジーで接続された仮想プロセッサの組とを備えている。仮想マシンは抽象物理的マシンにマッピングされ、その際、各仮想プロセッサは抽象物理的プロセッサにマッピングされる。抽象体の第3の層はスレ

ッド18である。これらのスレッドは、仮想プロセッサ上でランするライトウエイトのプロセスである。

【0060】仮想トポロジーは、例えば、メッシュ・トポロジーで物理的に接続された物理的プロセッサにマッピングされる仮想プロセッサのツリーである。仮想トポロジーによって、プログラマは、実施すべきアルゴリズムに適した(仮想)トポロジーでプログラムを表現することが可能となる。スティングは、仮想トポロジーから、ターゲット・マシンの実際の物理的トポロジーへの効率的なマッピングを提供する。また、仮想トポロジーによって、並列プログラムを、異なる物理的トポロジー間で容易に移すことが可能となる。

【0061】スティングの調整部分言語の主な構成要素は、ライトウエイト・スレッドと仮想プロセッサである。スレッドは、ローカル記憶装置(すなわち、レジスタ、スタック、ならびにヒープ)、コード、ならびに関連する状態情報(すなわち、ステータス、優先順位、プリエンブション・ビット、ロックなど)を含む単純なデータ構造である。それらは独立した制御の場所を定義している。このシステムは、スレッドが含むコードに対して制約を課さない。有効なスキームの表現はすべて、独立したプロセスとして扱われる。

【0062】図2、図3に示すように、各仮想プロセッサ(VP)16はスレッド・コントローラ(TC)17を含み、このコントローラはスレッド上およびスレッド・ポリシー・マネージャ(TPM)19上で状態遷移機能を実施する。そして、スレッド・ポリシー・マネージャはスレッドのスケジューリングと負荷平衡/移行ポリシーの両方を実施する。同じ仮想マシン内で各VPはスレッド・コントローラを共有するが、異なるVPは異なるスレッド・ポリシー・マネージャを持つことができる。

【0063】仮想プロセッサ16は物理的プロセッサ12上に、スレッド18が仮想プロセッサ上に多重化されているのと同じようにして多重化されている。各物理的プロセッサは、マルチプロセッサ環境における計算エンジンに対応している。各物理的プロセッサPPに関連しているのは仮想プロセッサ・コントローラ13および仮想プロセッサ・ポリシー・マネージャ15である。仮想プロセッサ・コントローラは、プリエンブションによって、あるいは明示的なリクエストによって、仮想プロセッサ間でコンテキスト・スイッチを行う。仮想プロセッサ・ポリシー・マネージャは、物理的プロセッサPP上で実行する仮想プロセッサ16に対するスケジューリングの決定を扱う。例えば、仮想プロセッサは、その上でスレッドが実行していない場合、そして他のVPからスレッドを移転できない場合には、物理的プロセッサの制御を放棄することができる。物理的プロセッサは、システム内のいかなる仮想マシンの仮想プロセッサをもランさせることができる。

【0064】仮想マシンは単一のアドレス空間24を含み、関連する仮想プロセッサはそれを排他的にアクセスすることができる。仮想マシンは、グローバル記憶プール26内のグローバルな情報（例えば、ライブラリ、ファイル・システムなど）を共有することができ、そしてグローバル共有オブジェクト28（すなわち、グローバル・アドレス空間にあるオブジェクト）をそれらのローカル・アドレス空間にマッピングする。仮想マシンはまた、アドレス空間内のすべての活性オブジェクトをトレースするために用いられる活性オブジェクト・グラフ（すなわち、ルート環境30）のルートを含んでいる。

【0065】すべてのスティング・オブジェクト（スレッド、VP、仮想マシンを含む）は持続性メモリ内に存在する。このメモリは個別領域の集合として構成されている。オブジェクトは、世代コレクタを用いて領域内に集められたガーベッジである。1つのオブジェクトはそのアドレス空間内の他のオブジェクトをすべて参照することができる。最初、オブジェクトは短寿命のスレッド・ローカル領域に存在する。ガーベッジ・コレクションから生き残ったオブジェクトは世代階層において上位に移る。この機能はユーザにとっては全く明らかである。

【0066】ファーストクラスのオブジェクトは、プロシージャに対して引数として渡したり、結果としてプロシージャから戻したり、あるいはデータ構造内に記憶できるオブジェクトのことである。本発明の抽象物理的マシンの望ましい実施例では、抽象物理的プロセッサ、仮想マシン、仮想プロセッサ、スレッドのグループ、ならびにスレッドはすべてファーストクラスのオブジェクトである。他の実施例では、スレッドおよび仮想プロセッサのみがファーストクラスのオブジェクトである。

【0067】スティング・コンパイラはオービット（Orbit）の改良バージョンである。オービットについてはD. Kranzらの論文に記述されている（参考文献：“Orbit: An Optimizing Compiler for Scheme”, in ACM SIGPLAN Notices, 21 (7): 219-233, July 1986）。コンパイラにより見えるターゲット・マシンは、現在ランしているスレッド・オブジェクトに対する参照を保持する専用のスレッド・レジスタを含んでいる。さらに、レジスタをコンテキスト・スイッチ上で退避、復元したり、あるいはスレッドの記憶領域（すなわち、スタックおよびヒープ）を割り当てたりするといった時間的な制約の厳しいオペレーションは、基本オペレーションとして用意される。連続するスキーム・プログラムは変更無しにコンパイルし、実行される。スティングでは、フューチャ、分散データ構造、ならびにスペキュラティブ平行オペレーションも実現している。スキーム・プログラムは、これらのパラダイムのいずれかによってサポートされた平行オペレーションによって自由に拡大させることができる。

【0068】スレッドは、スティングにおけるファーストクラスのオブジェクトである。従って、それらは引数としてプロシージャに渡すことができ、また結果として戻し、さらにデータ構造内に格納することができる。スレッドは、スレッドを生成したオブジェクトより長く生き残ることができる。スレッドの状態は、サンク（thunk）、すなわちスレッドが実行されるとき発動されるヌラリー（nullary）プロシージャを含んでいる。アプリケーションの値は終了時にスレッド内に格納される。例えば、 $(fork-thread(+y(*x*z)))$ という表現を評価することによって、サンク $(lambda() (+y(*x*z)))$ を発動する制御のライトウエイト・スレッドが生成される。このサンクの評価環境は、 $fork-thread$ という表現の辞書的環境である。

【0069】スレッドは状態情報をその状態の一部として記録する（図4および図5参照）。スレッドは、遅延36、スケジュール38、評価40、吸収42、あるいは確定44のいずれかの状態をとる。遅延されたスレッドは、スレッドの値が明確に要求されない限り、ランされることはない。スケジュールされたスレッドは、いずれかのVPが知っているスレッドであるが、まだ記憶資源は割り当てられていない。評価を行っているスレッドはランし始めたスレッドである。スレッドは、そのサンクの実行アプリケーションが結果を出すまでこの状態に留まる。上記結果が出たときスレッドの状態が確定する。吸収されたスレッドは、評価中のスレッドを特別化したものであり、重要であるため、以下にさらに詳しく説明する。

【0070】状態情報および評価すべきコードに加えて、1つのスレッドはまた、（1）それが完了するのを待っている他のスレッドに対する参照情報と、（2）サンクの動的な、そして例外の環境と、スレッドの親、兄弟、ならびに子を含む系統情報とを含んでいる。

【0071】各スレッドも、流体（すなわち動的）結合および例外の扱いを実現するために用いる動的な、そして例外の環境を有している。系統情報は、デバッグとプロファイリングのツールとして有用であり、それによってアプリケーションはプロセス・ツリーの動的な展開をモニタすることが可能となる。

【0072】スレッドの実現においては、言語における他の基本オペレーションを変更する必要はない。スレッドの同期化意味規則は、例えばMultilispの“touch”や、Lindaのタプル空間や、CMLの“sync”によって利用できる同期化機能をより一般的な（低レベルではあっても）形にしたものである。

【0073】アプリケーションは状態を完全に制御し、その状態のもとで、ブロックされたスレッドを復活させることができる。しかし、データフロー（すなわちフューチャ・タッチ）、非決定論的な選択、ならびに制約

にもとづく同期化または障壁同期化に対する明示的なシステム・サポートがある。

【0074】ユーザは、スレッド・コントローラ (TC) (スレッドのある状態において同期状態の遷移を実現する) が定義する1組のプロシージャ (以下にリストアップする) によってスレッドを操作する。TCは、レジスタの退避および復元という2つの基本オペレーションを除いて、全体をスキームによって書くことが望ましい。スレッド・コントローラは記憶領域を割り当てない。従って、TCのコールはガーベッジ・コレクションをトリガーしない。これらのオペレーションに加えて、スレッドは、プリエンブションのため、コントローラに入ることができる。スレッド・プロシージャを以下に示す。

【0075】(fork-thread expr vp) は、expr を評価するためにスレッドを生成し、それをvp上でランするようにスケジュールする。

【0076】(dcaly-thread expr) は、(スレッド値によって) 要求されたときexprを評価する遅延されたスレッドを生成する。

【0077】(thread-run thread vp) は、遅延された、ブロックされた、あるいは保留されたthreadをvpのレディー待ち行列に挿入する。

【0078】(thread-wait thread) は、このオペレーションを実行しているスレッドに、threadの状態が確定するまでブロックさせる。

【0079】(thread-block thread . blocker) は、threadにブロックすることをリクエストする。blockerは、スレッドがブロックするときの条件である。

```
(let ((filter-list (list)))
  (sieve (lambda (thunk)
            (set filter-list (cons (delay-thread (thunk))
                                   (filter-list))))))
```

では、フィルタがレイジーに生成される。フィルタは、一度要求されると、反復的に入力ストリームから要素を除去し、そして潜在的な素数を出カストリーム上に発生する。ラウンド・ロビン・スレッド配置規律を用いるV※

```
(thread-run (car filter-list)
             (mod (1+ (vm. vp-vector (vp. v
                                         m (current-virtual-processor))))
                   n))
```

(vp. vm (current-virtual-processor)) という表現は、現在のVPを一部とする仮想マシンを定義している。仮想マシンのパブリック・ステートは、その仮想プロセッサを収容するベクトル

* 【0080】(thread-suspend thread . quantum) は、スレッドに実行の保留をリクエストする。quantum引数が与えられた場合には、指定された期間が経過したときスレッドは再開される。そうでない場合には、スレッドは、thread-runを用いて明示的に再開されるまで、無期限に保留される。

【0081】(thread-terminate thread . values) は、threadに対してvaluesをその結果として終了することをリクエストする。(yield-processor) は、現在のスレッドに、そのVPの制御をやめるようリクエストする。このスレッドは適切なレディー待ち行列に挿入される。

【0082】(current-thread) は、このオペレーションを実行しているスレッドを復帰する。

【0083】(current-virtual-processor) は、このオペレーションが、その上で評価されている仮想プロセッサを復帰される。

【0084】ユーザがいかにスレッドをプログラムできるかを説明するため、図6のプログラムについて考える。このプログラムは、簡単な素数発見手段の実現を定義したものである。この定義ではいかなる特定の並行パラダイムも参照していない。このような問題はそのop引数によって抽象される。

【0085】この素数発見手段の実現は、ストリーム・アクセスにおけるブロッキング・オペレーション (hd)、およびスレッドの最後に付加するアトミック・オペレーション (attach) を与える、ユーザが定義した同期スレッド抽象体に依存している。

【0086】非同期の振舞の程度が異なる、素数発見手段の種々の処理を定義できる。例えば、

※P上でスケジュールした新しいフィルタを始動させるため、次のように書くことができる。

【0087】

ルを含んでいる。

【0088】シーブに対する上記コールを少し書き直すことにより、よりレイジーな素数発見手段の実現を表現できる。

【0089】

```

(let ((filter-list (list)))
  (sieve (lambda (thunk)
    (set filter-list
      (cons (create-thread
        (begin
          (map thread-block filter
            (thunk)))
          filter-list))
      (map thread-block filter
        -list)))
    n))

```

この定義では、潜在的な素数 p に遭遇したフィルタは、レイジーなスレッド・オブジェクト l を生成し、チェーン内の他のすべてのフィルタにブロックすることをリクエストする。 l の値が要求されたときは、フィルタはチェーン内のすべての要素をアンロックし、そしてその入力スレッドにおける p のすべての倍数を取り除く。この*

```

(sieve (lambda (thunk)
  (fork-thread (thunk) (current-vp)))
  n)

```

このアプリケーションを評価することによって、素数のすべての倍数を取り除くための新たなスレッドがスケジューされる。このスレッドは、この操作を実行する仮想プロセッサ上でスケジューされる。このコールでは、素数が新たに見つかるごとに、評価するスレッドが発生される。

【0092】スティングでは、スレッドのオペレーションを通常のプロシージャとして扱い、スレッドのオペレーションで参照されるオブジェクトを、スキームのどれか他のオブジェクトとして操作する。共通のストリームによって結ばれた2つのフィルタが終了した場合、上記ストリームが占有する記憶領域は再利用することができる。スティングは、スレッドのアクセスに対して先験的な同期化プロトコルを課さない。アプリケーション・プログラムが、スレッドの調整を整える抽象体を構築するようにしている。

【0093】フィルタによって生成されたスレッドは2つの方法の中の1つによって終了される。シーブに対するトップレベルのコールは、それがこれらのスレッドに対して明示的なハンドルを有するように、構成することができる。レイジーなシーブを生成するために用いるフィルタ・リスト・データ構造はその一例である。次に、

```

(map thread-terminate filter-list)

```

を評価して、シーブ内のすべてのスレッドを終了させることができる。あるいは、アプリケーションはスレッドのグループを用いて、これらのスレッドを集散的に管理

*コールでは要求にもとづいて、シーブの拡張および入力の消費を抑制する。

【0090】このシーブは次のように、よりイーガーなバージョンに変えることもできる。

【0091】

することができる。

【0094】<スレッド・グループ>スティングは、関連するスレッドの集まりに対する制御を獲得する手段としてスレッド・グループを与える。1つのスレッド・グループは、fork-thread-groupに対するコールによって生成される。このオペレーションは、新しいグループおよび新しいスレッドを生成し、新しいスレッドは新しいグループのルート・スレッドになる。子スレッドは、新しいグループを明示的に生成しない限り、同一のグループを、その親として共有する。1つのグループは1つの共有ヒープを含み、そのメンバーはすべてこのヒープをアクセスできる。スレッド・グループが次のコールによって終了したとき、

```

(thread-group-terminate group)

```

グループ内の生きているスレッドはすべて終了され、その共有ヒープはガーベッジ・コレクトされる。

【0095】スレッド・グループはまた、そのメンバーに対して、それをすべてひとまとめにして適用できるデバッグ・オペレーションおよびスレッド・オペレーションも含んでいる。スレッド・グループは、デバッグおよびモニタのためのオペレーション（例えば、与えられたグループ内のすべてのスレッドのリスタアップ、すべてのグループのリスタアップ、プロファイリング、系統の報告など）と共に、通常のスレッドのオペレーション（例えば、終了、保留など）と同種のオペレーションを提供する。従って、スレッド T が終了したとき、ユーザ

はTのすべての子(終了されるべきTのグループの一部として定義されている)に対して次のようにリクエストできる。

【0096】(thread-group-terminate(thread, group T))

スレッド・グループは、階層的メモリ・アーキテクチャにおいて、共有を制御するための重要なツールである。グループのメンバーが共有するオブジェクトは、グループの共有ヒープ内に含まれているので、これらオブジェクトはメモリ内で物理的に互いに近接していることが望ましく、それによってより良いローカリティが得られる。スレッド・グループはまた、スケジューリングの場として用いることもできる。例えば、スレッド・ポリシー・マネージャは、グループ内のすべてのスレッドがランすることを許可されない限り、グループ内のスレッドはいずれもランできないというスケジューリング・ポリシーを実現できよう。このスケジューリング方式は“ギャング・スケジューリング”プロトコルと同種のものである。スレッド・グループはデータのローカリティを改善するために仮想トポロジーと共に用いることができる。

【0097】<実行コンテキストおよびスレッド制御ブロック>スレッドが評価を開始したとき、実行コンテキストがそれに対して割り当てられる。評価を行っているスレッドはいずれも、スレッド制御ブロック(TCB)32(図5)としても知られる実行コンテキストと関連している。TCBはコンティニューエーションを一般的に表したものであり、それ自身のスタック31とローカル・ヒープ33を含んでいる。スタックとヒープはともに拘束でき、そしてヒープは生成スキャベンジング・コレクションを用いてガーベッジ・コレクションされる。記憶オブジェクト以外に、TCBは関連するロックと、スレッドが最後にコンテキスト・スイッチを実行したとき残っている、生きたレジスタすべての値と、スレッドのサブステート(例えば、初期化、レディー、評価、ブロック、保留などの状態)と、スレッドが最後に実行されたVPと、スレッドの優先順位と、タイム・クオンタムとを含んでいる。

【0098】スレッド・ステートおよびスレッド・サブステートの遷移図を図4に示す。TCBの状態は、評価を行っているスレッド上で許可されたオペレーションを反映している。評価中のスレッドTがTCB T_{TCB}を有しているなら、T_{TCB}のステート・フィールドは以下の中のいずれか1つを示す。

【0099】初期化46:T_{TCB}に関連するスタックとヒープが初期化されているが、どのコードもまだ実行されていない。

【0100】レディー48:Tは利用できるいかなるVP上でも実行できるが、いずれのVP上でも現在、まだ実行されていない。

【0101】ラン50:TはあるVP上で現在実行されている。

【0102】ブロック52:Tは、あるスレッド上で、またはある条件のもとで現在ブロックされている。

【0103】保留54:Tは、基本的に無期限に保留されている。

【0104】終了56:Tは実行を終了し、残りの状態を一掃している。

【0105】スレッドとは異なり、TCBはファーストクラスの、ユーザに見えるオブジェクトではない。スレッド・コントローラとスレッド・ポリシー・マネージャのみがそれらをアクセスできる。新しいスレッドがランのレディー状態にあるとき、TCBはそれに割り当てられる。スレッドが確定状態となったとき、スレッド・コントローラはそのTCBを、後に生成されるスレッドのために、利用できる。TCBはユーザが維持するデータ構造内に逃げ込むことはない。TCBはシステム・レベルのプロシージャによって排他的に操作される。

【0106】スティングの実現はスレッドに対する記憶領域の割り当てを必要とするまで延期する。他のスレッド・パッケージでは、スレッドを生成する動作は、単にフォークされるべきスレッドに対する環境を設定するだけでなく、記憶領域の割り当ておよび初期化も含んでいる。このアプローチでは2つの重要な点で効率の低下を招く。第1に、グラニュラリティの細かい並列のもとでは、スレッド・コントローラは、実際にスレッドをランさせることより、それらを生成し、初期化することにより長い時間を消費する。第2に、スタックおよびプロセス制御ブロックはスレッドが生成されると直ちに割り当てられるので、スレッド間のコンテキスト・スイッチはしばしば、キャッシュとページのローカリティの利点を活用できない。さらに、TCBの割り当てが遅延されない場合には、システムに必要な全メモリ容量は大幅に増加することになる。

【0107】スティングのスレッド制御ブロックは、仮想プロセッサによって管理されるリサイクル可能な資源である。TCBは、スレッドが評価を開始したときのみスレッドに対して割り当てられる。この割り当てのストラテジーはデータのローカリティを改善するように設計されている。TCBは、VP V上でランするべきスレッドTに対して4つの方法の中の1つによって割り当てることができる。

【0108】1. 現在V上で実行中のスレッドが終了した場合には、そのコンテキストは直ちに再割り当てのために利用できる。そのTCBは割り当てのための最も良い候補である。なぜなら、このTCBは、そのVPに対して最も高いローカリティを有しているからである。このVPに関連する物理的キャッシュおよびメモリは、最も最近VP上でランしたスレッドの実行コンテキストを含んでいる可能性が最も高い。

【0109】2. 現在実行しているスレッドが終了していない場合には、Tに対するTCBは、V上に維持されているTCBのLIFOプールから割り当てられる。ここでも再び、上記実行コンテキストが、最も高いローカルリティを有するものとなっている。

【0110】3. Vのプールが空の場合には、新しいTCBが、これもLIFOの順序で構成されたグローバル・プールから割り当てられる。ローカルVPプールはいずれも、それが保持できるTCBの数のしきい値 τ を維持している。プールがオーバーフローした場合には、そのVPは、ローカル・プール内のTCBの半分をグローバル・プールに移動する。ローカル・プールがオーバーフローしていない場合には、 $\tau/2$ TCBがグローバル・プールからVPのローカル・プールに移動される。グローバル・プールは2つの役割を果たす。すなわち、

(1) TCBの割り当ておよび再使用におけるプログラムの振舞の影響を最小化すること、および(2)すべての仮想プロセッサに対するTCBの公正な分配を保証することである。

【0111】4. 最後に、TCBをグローバル・プールあるいはローカル・プールのいずれにおいても利用できない場合には、 $\tau/2$ TCBの新しい組が動的に生成され、Tに割り当てられる。新しいTCBは、グローバル・プールおよびVPのローカル・プールがともに空の場合にのみ生成されるので、スティング・プログラムの評価の際に実際に生成されるTCBの数は、すべてのVPによって集散的に決められる。

【0112】＜仮想プロセッサ＞仮想プロセッサ（拡張して、仮想マシン）は、スティングではファーストクラスのオブジェクトである。ファーストクラスというVPの状態には、スティングを高レベルのスレッド・システムおよび他の非同期並列言語のいずれからも区別する重要な意味がある。第1に、明示的にプロセスを特定の仮想プロセッサにマッピングすることによって並列計算を組織できる。例えば、VP V上で実行している他のスレッドQと密接に通信することが知られているスレッドPは、トポロジ的にVに近いVP上で実行すべきであ*

(define (up-VP)

(let ((address (vp-address (current
ly-virtual-processor))))

(array-ref 3D-mesh (vector-ref
address 0))))

1. 適当な物理的プロセッサにマッピングされる仮想プロセッサの組を生成する。

【0116】2. 仮想トポロジーにおけるアドレスを各VPに関連づける。

【0117】3. 仮想トポロジーにおいて絶対アドレッシングのために用いるデータ構造に仮想プロセッサを格納し、その構造上に適切なアクセスルーチンを定義する。

【0118】4. 自己相対アドレッシングのプロシージャ 50

*る。スティングでは、VPは直接的に示されるので、このような考慮を実現することができる。例えばシストリック・スタイルのプログラムは、現在のVP（例えば、現在VP、左VP、右VP、上VPなど）から離れて自己相対アドレッシングを用いて表現することができる。このシステムは、多数の共通トポロジー（例えば、ハイパーキューブ、メッシュ、シストリック・アレーなど）に対していくつかのデフォルト・アドレッシング・モードを提供する。さらに、VPは特定の物理的プロセッサにマッピングできるので、ファーストクラスのデータ値として仮想プロセッサを操作できるという能力により、スティングのプログラムは、種々の特定のプロセッサ・トポロジーで定義される異なる並列アルゴリズムを極めて柔軟に表現することができる。

【0113】図7のプログラムを参照して説明する。このプログラムは、物理的プロセッサの2次元メッシュ上で多重化された仮想プロセッサの3次元メッシュを生成するものである。このアレーは、物理的マシンの高さおよび幅と同じ高さおよび幅を有している。深さ方向の各要素を同じ仮想プロセッサにマッピングすることによって、3次元アレーを2次元アレーに縮小する。従って、生成された仮想プロセッサの数は、物理的プロセッサの数と同じである。同じ深さのプロセッサにマッピングされたスレッドはすべて同じVP上で実行する。プロシージャget-pm-heightとget-pm-widthは物理的マシン・インターフェースによって与えられる。仮想プロセッサの絶対アドレッシングは、create-3D-meshだけ戻したアレーへの単純なアレー参照である。

【0114】create-vpプロシージャは、get-ppが戻した物理的プロセッサ上で走る新しいVPを生成する。トポロジーが生成されると、現在のVPから離れて自己相対アドレッシング・プロシージャを構築することが可能である。例えば、トポロジーにおいて1ディメンジョン上方に移動する上VPプロシージャを定義することができる。

【0115】

を定義する。

【0119】＜スレッド・コントローラ＞スレッド・コントローラは、仮想プロセッサによる、物理的プロセッサやスレッドなど、他のシステム要素とのやり取りを扱う。スレッド・コントローラの最も重要な機能は、スレッドの状態遷移を扱うことである。スレッドが、その状態遷移によって、現在その上でランしている仮想プロセッサを生じた場合には、必ずスレッド・コントローラは

スレッド・ポリシー・マネージャをコールし、次にどのスレッドをランするべきかを定める。

【0120】スティングのスレッド・コントローラを実現する場合、いくつかの興味深い問題が明らかになる。中心的な状態遷移プロシージャは図9および図10に示す。これらのプロシージャで見られるTCBでの操作は、ユーザ・アプリケーションでは利用できない。スレッド・コントローラはスティングの中にかかれていて、TCBプロシージャに対するすべての同期コールは通常のプロシージャ・コールとして扱われる。従って、現在のスレッドでランしているプロシージャが用いる活性レジスタは、コントローラへのエントリのとき、スレッドのTCB内に自動的に退避される。

【0121】プロシージャstart-context-switch (図8)は、その引数として、現在のスレッド (すなわち、TCに入ったスレッド) に対する望ましい次の状態をとる。プリエンブションは最初にディスエーブルされる。次に、新しいスレッド (あるいはTCB) が、プロシージャtpm-get-next-threadによって復帰される。

【0122】ランできるスレッドが無い場合には、プロシージャは偽 (false) を返す。この場合、現在のスレッドは再度ランされるか (レディー状態にあるとして)、あるいはプロシージャtpm-vp-idleが、現在のVPを引数としてコールされる。プロシージャtpm-vp-idleは種々の簿記操作を行うことができ、また、その物理的プロセッサに他のVPに切り換えるようリクエストすることができる。

【0123】次のオブジェクトが現在のTCBである場合、動作は一切行われず、現在のスレッドが直ちに再開される。戻されたオブジェクトが他のTCBの場合には、その状態がランに設定され、VPフィールドは現在のVPに設定される。そして、現在のTCBは (その状態がデッドの場合) TCBプール内でリサイクルされるか、またはそのレジスタが退避され、そして新しいTCBの状態が、プロセッサ・レジスタに還元される。

【0124】戻されたオブジェクトが、実行コンテキストを持たないスレッドの場合には、TCBがそれに対して割り当てられる。このTCBは、next-stateフィールドがデッドの場合には現在のTCBとなる。あるいはVPローカル・プールまたはグローバル・プールから割り当てられるTCBとなる。スレッドは、基本プロシージャstart-new-tcbを用いて実行を開始する。このスレッドは、その実行コンテキストとして新しいTCBを用い、プロシージャstart-new-thread (図10参照) を応用する。

【0125】finish-context-switchのコード (図9) は、start-context-switchが復帰させたスレッドによって実行される。その目的は、新しいスレッドのVPフィールドを設

定するためにスイッチ・アウトされたスレッド (このプロシージャ内で以前にコールされている) が保持するロックを解放し、適切であるなら以前のものをレディー待ち行列に組み入れ、プリエンブションタイムを再設定する。新しいスレッドがVP上に設定された後でのみ以前のものを待ち行列に組み入れることにより、コントローラは、状態遷移を起させることと、スレッドをVPのレディー待ち行列に組み入れることとの間の競合状態を排除する。プロシージャtmp-enqueue-ready-threadとtmp-enqueue-suspended-threadは、スレッド・ポリシー・マネージャによって実現される。

【0126】start-new-threadのコードを図10に示す。サンクE_iを有するスレッド・オブジェクトは、それに対してTCBが割り当てられると、評価を開始でき、そしてデフォルト・エラー・ハンドラーおよび適当なクリンアップ・コードと関連するようになる。E_iから出るためのスロー (start-new-threadが設定するキャッチポイント) はスレッド・スタックに適切に巻き戻させ、それによってスレッドが保持するロックなどの資源が適切に解放されるようにする。E_iの評価に続く退出のコードはスレッドのスタックとヒープをガーベッジ・コレクションし、E_iが生成した値をスレッド状態の一部として格納し、この値を待っているスレッドをすべて目覚めさせ、状態遷移プロシージャに対するテイル・リカーシブ・コールに、ランすべき新しいスレッドを選択させる。E_iはダイナミック・wind・フォーム内に含まれているので、スレッドが異常終了した場合でも、スレッドの記憶領域がガーベッジ・コレクションされることが保証される。

【0127】ガーベッジ・コレクションは、スレッドのウェイターが起される前に行われなければならない。それは、スレッド (スレッドのサンクが復帰させたオブジェクトを含む) より長生きであって、ローカル・ヒープを含んでいたオブジェクトは他の活性ヒープに移転させる必要があるためである。これが行われないと、他のスレッドが、新たに終了したスレッドの記憶領域に対する参照を得ることになるからである。確定したスレッドの記憶領域は他のスレッドに割り当てられるので、これは明らかにエラーとなる。

【0128】<スレッド・ポリシー・マネージャ>各仮想プロセッサはスレッド・ポリシー・マネージャを有している。スレッド・ポリシー・マネージャは、仮想プロセッサ上でのスレッドのスケジューリングおよび移行に関するすべてのポリシーの決定を行う。スレッド・コントローラはスレッド・ポリシー・マネージャの依頼者であり、ユーザのコードはそれをアクセスできない。スレッド・コントローラは、次のことに関連して決定を行う必要がある場合には必ずスレッド・ポリシー・マネージャをコールする。すなわち、スレッドの仮想プロセッサ

への初期マッピングと、現在のスレッドがなんらかの理由で仮想プロセッサを解放したとき、次に仮想プロセッサはどのスレッドをランさせるべきかということと、いつ、どのスレッドを仮想プロセッサに、あるいは仮想プロセッサから移転させるかということである。

【0129】すべての仮想プロセッサは同一のスレッド・コントローラを有しているが、各仮想プロセッサは異なるポリシー・マネージャを備えることができる。このことは、各プロセッサが、必要なスケジューリングがさまざまな異なるサブシステムを制御するというリアルタイム・アプリケーションにとって特に重要である。

【0130】スレッド・ポリシー・マネージャはスレッド・コントローラに対してよく定義されたインターフェースを提供する。スレッド・ポリシー・マネージャが決定を行うために用いるデータ構造は、スレッド・ポリシー・マネージャにとって完全にプライベートなものとなっている。それらは特定のスレッド・ポリシー・マネージャに対してローカルとしたり、あるいは種々のスレッド・ポリシー・マネージャが共有するようにでき、また、それらの組み合わせとすることもできる。しかし、システムの他の部分は一切利用できない。従って、スレッド・ポリシー・マネージャは、異なる仮想マシンに対して異なる振舞を行うようにカスタマイズすることができる。その結果、ユーザは、ランさせるプログラムの種類に応じてポリシーの決定をカスタマイズすることができる。

【0131】VPはそれぞれ異なるスレッド・ポリシー・マネージャを備えることができるので、アプリケーションによって生成された異なるグループのスレッドは、異なるスケジューリング方式の対象とすることができる。仮想マシンあるいは仮想プロセッサは異なるスケジューリング・プロトコルあるいは異なるスケジューリング・ポリシーを扱うよう調整することができる。

【0132】スティングのスレッド・コントローラは、スレッドの状態遷移プロシーダを定義するが、先験的なスケジューリング・ポリシーあるいは先験的な負荷平衡・ポリシーは定義しない。これらのポリシーはアプリケーションに依存する場合がある。いくつかのデフォルト・ポリシーがスティングの全実行時間環境の一部として与えられるが、ユーザは自身のポリシーを自由に書くことができる。事実、図3に示すように、各仮想プロセッサ16はそれ自身のスレッド・ポリシー・マネージャ(TPM)19を有している。従って、与えられた仮想マシン内の異なるVPは異なるポリシーを実現できる。TPM19はスレッドのスケジューリング、プロセッサ／スレッドのマッピング、ならびにスレッドの移行を扱う。

【0133】アプリケーションを個別のスケジューリング・グループに分けられるということは、長寿型の並列(あるいは会話型)プログラムにとって重要である。I

／Oに関連したプロシーダを実行するスレッドは、計算に関連したルーチンを実行するスレッドとは異なるスケジューリングを必要とする。リアルタイムの制約を持つアプリケーションは、単純なFIFOスケジューリング・ポリシーのみを必要とするものとは異なるスケジューリング・プロトコルを用いて実現されるべきである。

【0134】ツリー構造の並列プログラムは、LIFOにもとづくスケジューラを用いることによって、もっとも良好に動作しよう。マスタ／スレーブ・アルゴリズムあるいはワーカー・ファーム・アルゴリズムをランさせるアプリケーションは、公正さのためにラウンド・ロビン・プリエンブション・スケジューラを用いることによって、より良好に動作しよう。これらのアプリケーションはすべて、大きいプログラム構造体あるいは大きいプログラム環境の構成要素であるから、これらのアプリケーションを異なるポリシー・マネージャによって評価することで得られる柔軟性は重要である。同一の仮想マシン上で評価するスレッドの集まりを独立に実行する、個別のアプリケーションは存在し得る。さらに、各個別のスケジューラは、異なる性能特性を有し、そして異なる形で実現されたスレッド・ポリシー・マネージャを有することができる。

【0135】本発明は、柔軟なフレームワークの提供を探究するものである。そしてこの柔軟なフレームワークは、スレッド・コントローラ自身に対する変更を行うことなく、ユーザに対して明らかに異なるスケジューリング方式を組み入れることができるものである。そのため、すべてのTPMは、その実現において制約は一切課されていないが、同一のインターフェースに従わなければならない。以下に示すインターフェースは、ランすべき新しいスレッドを選択し、評価中のスレッドを待ち行列に挿入し、スレッドの優先順位を設定し、そしてスレッドを移行させるためのオペレーションを提供する。これらのプロシーダはTCが排他的に用いるためのものである。一般に、ユーザ・アプリケーションは、スレッド・ポリシー・マネージャとスレッド・コントローラとのインターフェースを承知している必要はない。

【0136】(tpm-get-next-thread vp)は次にvp上でランすべきレディー状態のスレッドを戻す。

【0137】(tpm-enqueue-ready-thread vp obj)は、スレッドあるいはTCBのいずれかであろうobjをvpに関連するTPMのレディー待ち行列に挿入する。

【0138】(tpm-priority priority)および(tpm-quantum quantum)は、それぞれの引数が有効な優先順位か、あるいは有効なクオンタムかを確認するガードプロシーダである。

【0139】(tpm-allocate-vp th

read) は thread を vp に割り当てる。vp が偽の場合には、thread は TPM によって確定される仮想プロセッサに割り当てられる。

【0140】(tmp-vp-idle vp) は、vp 上に評価を行っているスレッドが無い場合、スレッド・マネージャによってコールされる。このプロシージャはスレッドを他の仮想プロセッサから移行させたり、簿記を行ったり、他のVPに対するプロセッサ・スイッチ自身を持つために物理的プロセッサをコールしたりすることができる。

【0141】(tpm-enqueue-suspend-up-thread) は、vp の保留待ち行列上の thread を保留する。

【0142】TPM は、評価中のスレッドに対するスケジューリングの順序を決定する以外に、負荷平衡の2つの基本的決定を行う。(1) 新しく生成されたスレッドを走らせるべきVPを選択する。(2) VP上のどのスレッドを移行できるかを決め、他のVPからどのスレッドを移行させるかを決める。

【0143】最初の決定は、初期の負荷平衡を扱うために重要である。第2の決定は、動的負荷平衡・プロトコルをサポートするために重要である。新しく評価中のスレッドの最初の配置の決定は、しばしば現在評価中のスレッドの移行を決めるために用いられる優先順位とは異なる優先順位にもとづいて行われる。TPM インターフェースはこの区別を保存する。

【0144】スケジューリング・ポリシーはいくつかの重要な事柄に従って分類できる。

【0145】ローカルティ：このシステム内に単一のグローバル待ち行列があるか、あるいは各TPMはそれ自身の待ち行列を持っているか？

状態：スレッドはそれらの現在の状態にもとづいて区別されているか・例えば、あるアプリケーションは、すべてのスレッドが、それらの現在の状態に関係無く単一の待ち行列を占めるという実現法を選択するかもしれない。あるいは、スレッドが評価中か、スケジュールされたか、保留されているかなどにもとづいて、スレッドを異なる待ち行列に分類することを選択するかもしれない。

【0146】順序付け：待ち行列は、FIFO、LIFO、ラウンド・ロビン、優先順位、あるいはリアルタイムの構造体として(他のものの中で)実現されているか？

直列化：アプリケーションはどのようなロッキング構造を種々のポリシー・マネージャの待ち行列に課するか。

【0147】この分類でどの選択肢を選ぶかによって、結果としての性能特性に差が生じる。例えば、評価中のスレッド(すなわち、TCBを有するスレッド)をスケジュールされたスレッドから区別するグラニュラリティ構造体を適合させ、そしてスケジュールされたスレッド

のみを移行させることができるという制約を課した場合、評価中のスレッドの待ち行列をアクセスするのにロックは不要となる。この待ち行列は、それが生成されたVPに対してローカルである。しかし、スケジュールされたスレッドを保持している待ち行列は、他のVP上のTPMによる移行のターゲットであるから、ロックされなければならない。この種のスケジューリング方式は、動的負荷平衡が問題ではない場合には、有用である。従って、多数の長寿命の、非ブロッキング・スレッド(継続時間はほぼ同じ)が存在するときは、ほとんどのVPは、それら自身のローカル・レディー待ち行列上のスレッドの実行に、ほとんどの時間、ビジーとなる。従って、このようなアプリケーションにおけるこの待ち行列上のロックを除去することは有益である。一方、継続時間が変動するスレッドを発生するアプリケーションは、スケジュールされたスレッドおよび評価中のスレッドの両方の移行が可能なTPMと共に用いたとき、ラン可能なレディー待ち行列をロックすることに伴ってコストがかかるが、より高いパフォーマンスを示す。

【0148】スレッド・ポリシー・マネージャが新しいスレッドを実行する必要があるときは常に、グローバル待ち行列はスレッド・ポリシー・マネージャ間の競合を意味する。しかし、このような仕組みにすると、多くの並列アルゴリズムの実現において有用である。例えば、マスタ/スレーブ(あるいはワーカー・ファーム)プログラムでは、マスタに最初にスレッドのブールを生成する。これらのスレッドは、それら自身はいかなる新しいスレッドも生まない、長寿命の構造体である。これらは一度VP上でランすれば、滅多にブロックすることはない。従って、このようなスレッドを実行しているTPMは、ローカル・スレッド待ち行列を維持することのオーバーヘッドをサポートする必要はない。しかし、ローカル待ち行列は、プロセスの構造がツリーあるいはグラフの形をとる結果、並列プログラムの実現においては有用である。これらの待ち行列は、仮想プロセッサの組において公正にスレッドをロード・バランスするために、このようなアプリケーションで用いることができる。

【0149】<メッセージ伝達抽象体>メッセージ伝達は分離メモリ・アーキテクチャにおいて効率の良い通信メカニズムでなければならない。特に、グラニュラリティの粗い並列アプリケーション、あるいは既知の通信パターンを有する並列アプリケーションに対してそうである。ポートは、分離メモリ・アーキテクチャ上で共有メモリを実現することのオーバーヘッドを最小限のものとするためにスティング内に設けられたデータ抽象体である。ファーストクラスのプロシージャおよびポートは、このコンテキストにおいて共同作業を示す。

【0150】スティングは、メッセージ伝達抽象体を共有メモリ環境において統合することを可能とする。ポートはファーストクラスのデータ・オブジェクトであり、

他のスレッドから送られるメッセージに対するレセプタクルとして働く。スティングは共有仮想メモリ・モデルを用いるので、いかなる複合データ構造（閉包を含む）でもポートを通じてやり取りできる。この柔軟性のため、スティングのアプリケーションはユーザ・レベルのメッセージ伝達プロトコルを明瞭な形で実現でき、そして単一化した環境において共有メモリとメッセージ伝達の最も優れた長所を結合することが可能となる。

【0151】ポートはファーストクラスのデータ構造である。ポートに対しては2つの基本的オペレーションがある。

【0152】1. (put obj port) は、obj を port にコピーする。この操作は送り手と非同期である。

【0153】2. (get port) は、port 内の最初のメッセージを除去し、port が空の場合にはブロックする。

【0154】ポートPから読み出したオブジェクトは、Pに書き込まれたオブジェクトのコピーである。このコピーは浅いコピーである。すなわち、オブジェクトの最上位の構造体のみがコピーされており、下位の構造体は*

```
(define (receiver port)
  (let ((msg (get port)))
    (fork-thread (msg) (current-vp))
    (receiver))))
```

送出されたプロシージャはこのレシーバの仮想プロセッサ上で評価される。レシーバは、メッセージを評価するために新しいスレッドを生成することによって、古いリクエストの処理と並行して新しいリクエストを受け入れることができる。

【0157】このスタイルの通信は“アクティブ・メッセージ”と呼ばれている。それは、メッセージを受け取ったとき行うべき動作が、基本のシステムの一部としてコード化されておらず、メッセージそれ自身によって決められているからである。仮想プロセッサとスレッドのインターフェースは、メッセージ通信をサポートするためにいかなる変更も必要としないので、このようなモデルによって極めて大きい柔軟性と単純性が得られる。スティングの設計における2つのことが、この機能の実現にとって重要である。(1) オブジェクトが共有仮想メモリに存在するため、すべてのオブジェクト（他のオブジェクトに対するレファレンスを有しているオブジェクト、例えば閉包を含む）は仮想プロセッサ間で自由に送信できる。(2) ファーストクラスのプロシージャは、ユーザが定義する複雑なメッセージ・ハンドラーの構築を可能とする。これらのハンドラーはいずれかの仮想プロセッサ上の分離したスレッド内で実行できる。分離メモリ・マシンでは、オブジェクトは分散共有仮想メモリに存在することになる。説明のため、上述の例で、Eをデータベースの複雑な問い合わせとする。このデータ

*共有されている。これらのポートは、共有メモリが不十分な場合に用いよう設計されているので、意味規則をコピーすることで設計されている。put の標準バージョンはシャローコピーを行うが、ディープコピーを行うバージョンもある。そのバージョンは、最上位のオブジェクトをコピーするだけでなく、下位の構造体もすべてコピーする。

【0155】例えば、浅いコピーを用いてメッセージ内の閉包を送る場合、閉包のコピーを構築する。しかし、閉包が定義する環境内で束ねられたオブジェクトへの参照は保存する。使用するコピー・メカニズムの選択は、明らかに背後の物理的アーキテクチャとアプリケーションの分野の影響を受ける。スティング実現が存在する特定の物理的サブストレートに適合させることのできる一連のメッセージ伝達実現が存在する。

【0156】従って、つぎの表現の評価により、

```
(put (lambda () E) port)
プロシージャ (lambda () E) の閉包が port
へ送出される。port 上でレシーバが次のように定義
されているなら、
```

ベースが存在するプロセッサ上でレシーバが例示されたとすると、このような問い合わせは、データベース自身の、コストのかかる移行を伴わない。問い合わせは、データベースが存在するプロセッサに直接コピーされるので、通信のコストが低減される。データベースそれ自身は問い合わせを実行するプロセッサに移行する必要がない。より伝統的なRPCスタイルの通信ではなくデータにプロシージャを送るという能力により、いくつかの点で重要なパフォーマンスおよび表現性の向上が得られる可能性がある。

【0158】ファーストクラスのプロシージャおよびライトウエイトのスレッドは、アクティブ・メッセージにおいて、魅力的な高レベルの通信抽象体を伝達する。これらの抽象体を利用せずにアクティブ・メッセージをサポートするシステムでは、この機能は典型的には低レベル・サポート・プロトコルによって実現される。ファーストクラスのプロシージャはアクティブ・メッセージを平凡に実現することを可能とする。アクティブ・メッセージはポートに送られるプロシージャである。ファーストクラスのポートは分散計算環境においても明確で重要な効用を有し、そして従来のPRCより簡単で、かつ清潔なプログラミング・モデルの実現を可能とする。

【0159】＜メモリ管理＞スティングは共有仮想メモリ・モデルを用いる。分散メモリ・プラットフォーム上ではスティングは分散共有仮想メモリ・サブストレート

上で構築されなければならない。従って、参照の意味は、参照がどこで発生されているか、あるいはオブジェクトが物理的にどこにあるか、には依存しない。

【0160】＜記憶機構＞スティングでは各TCB32に関連して3つの記憶領域がある(図5)。第1はスタック31であり、スレッドによって生成されたオブジェクトの割り当てに用いられる。このスレッドの寿命は、それを生成したものの動的な範囲を越えない。より正確には、スタック上に割り当てられたオブジェクトは、現在の(あるいは前の)スタック・フレームに割り当てられた他のオブジェクト、あるいはヒープに割り当てられた他のオブジェクトしか参照できない。スタックが割り当てられたオブジェクトはヒープ内のオブジェクトを参照することができる。なぜなら、そのスタックに関連するスレッドは、ヒープ33がガーベッジ・コレクションされる間、保留となるからである。スタックに含まれている参照情報は、ガーベッジ・コレクタによってトレースしたとされるルートの一部である。

【0161】スレッドにとってプライベートなヒープ、すなわちローカル・ヒープ33は、割り当てられた非共有オブジェクトに対して用いられる。このオブジェクトは、その寿命が、オブジェクトを生成したプロセスの寿命を越える可能性がある。越える可能性があるとしたのは、スキームやMLなどのプログラミング言語ではコンパイラがオブジェクトの寿命を常に決めることができるとは限らないからである。さらに、未知のプロセスに対するコールが可能な言語においては、オブジェクトの寿命が決められない場合もある。プライベート・ヒープに含まれている参照情報は同じプライベート・ヒープ内の他のオブジェクト、あるいは共有ヒープ、すなわちグローバル・ヒープ35を示すことができるが、スタック31内のオブジェクトを示すことはできない。このスタック内の参照情報はプライベート・ヒープ内のオブジェクトを示すことができるが、共有ヒープ内の参照情報はこれらのオブジェクトを示せない。プライベート・ヒープに割り当てられたデータは、単一の、制御のスレッドによって排他的に用いられるので、プライベート・ヒープによってより高いローカルリティが実現する。複数のスレッド間にさしはさまれた割り当てがないという

ことは、ヒープ内で互いに接近したオブジェクトは、論理的に互いに関連している可能性が高いことを意味する。

【0162】他のスレッドは、スレッドのスタックあるいはローカル・ヒープ内に含まれているオブジェクトをアクセスできない。従って、スレッドのスタックおよびローカル・ヒープは共に、同期化あるいはメモリのコピーレンシーを考慮することなく、プロセッサ上のローカル・メモリにおいて実現することができる。スレッドのローカル・ヒープは実際には、世代的に組織した一連のヒープである。記憶領域の割り当ては常に、他の世代的

コレクタと同様に、最も若い世代において行われる。オブジェクトは、年齢が高くなるにつれて、古い世代に移動される。ローカル・ヒープのガーベッジ・コレクションはすべてスレッドそれ自身によって行われる。ガーベッジ・コレクションをサポートするほとんどのスレッド・システムでは、システム内のスレッドはすべて、ガーベッジ・コレクションの間は保留されなければならない。それに対して、スティングのスレッドは、他のスレッドと独立して、そして非同期的にそれらのローカル・ヒープをガーベッジ・コレクションする。従って、他のスレッドは、特定のスレッドがそのローカル・ヒープをコレクションしている間、計算を続けることができる。その結果、より優れた負荷平衡と高いスループットが得られる。このガーベッジ・コレクションの戦略の第2の長所は、ローカル・ヒープのガーベッジ・コレクションにかかるコストが、システム内のすべてのスレッドに課されるのではなく、記憶領域を割り当てるスレッドにのみ課されるという点にある。

【0163】スティングは、関連するスレッドの集まりを制御するための手段として“スレッド・グループ”を与える。子スレッドは、それが新しいグループの一部として生成されたのでない限り、その親と同一のグループに属する。スレッド・グループは、デバッグおよびモニタのためのオペレーション(例えば、与えられたグループ内のすべてのスレッドのリストアップ、すべてのグループのリストアップ、プロファイリング、システムの報告など)と共に、通常のスレッドのオペレーション(例えば、終了、保留など)を与える。さらに、スレッド・グループはまた、そのメンバーがすべてアクセスできる“共有ヒープ”を含んでいる。

【0164】スレッド・グループの共有ヒープ、すなわちグローバル・ヒープ35は、スレッド・グループが生成されたとき割り当てられる。ローカル・ヒープのような共有ヒープは実際には、世代的に組織された一連のヒープである。共有ヒープ内の参照情報は共有ヒープ内のオブジェクトしか示せない。これは、共有オブジェクトから参照されるオブジェクトはすべて共有オブジェクトであり、従って、共有ヒープ内に存在しなければならないからである。この共有ヒープに対する制約は、(a) 共有ヒープ内にあるか、あるいは(b) ローカル・ヒープ内に割り当てられていて、共有ヒープ内にガーベッジ・コレクションされているオブジェクトを、共有ヒープに記憶された参照情報が指示することを保証することによって、実施される。すなわち、参照されたオブジェクトから到達可能なオブジェクトのグラフは、共有ヒープ内にコピー、または配置されなければならない。このメモリ・モデルのオーバーヘッドは、ローカル・ヒープ上に割り当てられたオブジェクトに対する参照情報がどれくらい頻繁にエスケープするかによって決まる。経験によれば、ファイナライズ並列プログラムを実現す

る場合、ローカル・ヒープに割り当てられたオブジェクトはほとんど、関連するスレッドに対してローカルであり続け、共有されない。スレッド間で頻繁に共有されるオブジェクトは、言語抽象体あるいはコンパイル時の分析によって容易に検出される。

【0165】要約すると、あるスレッドに関連するスレッド領域間の参照規律は次のようになる。すなわち、

(1) スタック内の参照情報は、その現在のあるいは以前のスタック・フレーム、またはローカル・ヒープ、または共有ヒープ内のオブジェクトを示し、(2) ローカル・ヒープ内の参照情報は、そのヒープ上のオブジェクトあるいはなんらかの共有ヒープに割り当てられたオブジェクトを示し、そして(3) 共有ヒープ内の参照情報は、その共有ヒープ(あるいは、他のなんらかの共有ヒープ)に割り当てられたオブジェクトを示す。

【0166】ローカル・ヒープのように、グローバル・ヒープは世代的に組織されているが、グローバル・ヒープのガーベッジ・コレクションは、ローカル・ヒープに対するものより複雑である。それは、多数の異なるスレッドが、グローバル・ヒープ内のオブジェクトを同時にアクセスする場合があるからである。なお、その結果、グローバル・ヒープの割り当てにはヒープのロックが必要である。

【0167】グローバル・ヒープをガーベッジ・コレクションするために、関連するスレッド・グループ内のすべてのスレッド(そして、その下位のもの)は保留される。それは、これらのスレッドはすべてグローバル・ヒープ内のデータをアクセスできるからである。しかし、システム内の他のスレッド、すなわち、ガーベッジ・コレクションされるヒープに関連するグループ内にないものは、ガーベッジ・コレクションと無関係に実行を継続する。

【0168】各グローバル・ヒープは、それに関連し、そこに到来する参照情報を有している。これらの組は、領域境界を横断する参照情報の記憶に対するチェックによって、維持される。グローバル・ヒープに関連するスレッドが保留された後、ガーベッジ・コレクタは到来参照情報の組をガーベッジ・コレクションのためのルートとして用いる。到来参照情報の組から到達できるオブジェクトはすべて新しいヒープにコピーされる。ガーベッジ・コレクションが終了すると、グローバル・ヒープに関連したスレッドは再開される。

【0169】<抽象物理的マシンおよび抽象物理のプロセッサ>このオペレーティング・システムの最も低レベルの抽象体は、抽象物理的マシン(APM)と呼ばれるマイクロ・カーネルである。

【0170】APMはスティング・ソフトウェア・アーキテクチャにおいて3つの重要な役割を果たす。

【0171】1. 複数の仮想マシンをサポートする安全で効率的な基礎を提供する。

【0172】2. システム内の他のすべての要素を、ハードウェアに依存する特徴および特異性から分離する。

【0173】3. システムの物理的ハードウェアに対するアクセスを制御する。

【0174】APMはルート仮想マシンと呼ばれる特別の仮想マシン内で実現される。このマシンは、仮想アドレス空間、仮想プロセッサ、ならびにスレッドを含む、他のいずれの仮想マシンでも利用できる機能に対するアクセス手段を有している。さらに、ルート仮想マシンは、抽象物理的プロセッサ、デバイス・ドライバ、ならびに仮想メモリ・マネージャに対するアクセス手段を有している。抽象物理的マシンは仮想マシンによって構成されており、その結果、いくつかの重要な表現性が得られる。ヘビーウエイトのスレッドは一切無い。すべてのスレッドはライトウエイトである。システム・コールを実現するカーネル・スレッドあるいはスタックは無い。すべてのシステム・コールは、システム・コールを作成するスレッドの実行コンテキストを用いて扱われる。このことは、スキームが安全な言語であり(すなわち、ダングリング・ポインタ、アドレスとデータ間の自由強制などは不可能である)、そしてAPMの部分はシステム内のすべての仮想マシンにマッピングされているため、可能となっている。ユーザのスレッドが利用できる非同期的プログラミング構築体は、APM内のスレッドも利用できる。APMに関連したスレッドは、仮想マシン内の他のすべてのスレッドと同様に制御することができる。カーネル操作の実行をブロックするスレッドは、そのことを、それらスレッドの仮想プロセッサに通知する。それによってVPは他のなんらかのスレッドを自由に実行できる。これはスレッド間の通信およびI/Oの両方の場合に行われる。スティングは、例えば、スケジューラの起動、あるいはPsycheの仮想プロセッサ抽象体と同じ能力を提供するように、非ブロッキング・カーネル・コールを処理する。

【0175】仮想マシンはAPMによって生成され、そして破壊される。新しい仮想マシンの生成には以下のことが伴う。

【0176】1. 新しい仮想アドレス空間を生成する。

【0177】2. このアドレス空間にAPMカーネルをマッピングする。

【0178】3. この仮想マッピング内にルート仮想プロセッサを生成する。

【0179】4. このマッピングに抽象物理的プロセッサを割り当てる。

【0180】5. 抽象物理的プロセッサ上でランさせるために上記ルート仮想プロセッサをスケジュールする。

【0181】仮想マシンの破壊には、そのマシン上でランしているすべてのスレッドを終了させるための信号を発生し、マシン内で実行しているスレッドがオープンしたデバイスをすべてクローズし、そして最後に、このマ

シンに関連する仮想アドレス空間の割り当てを解除することが伴う。

【0182】各プロセッサ抽象体12は仮想プロセッサ・コントローラ(VPC)13と仮想プロセッサ・ポリシー・マネージャ(VPPM)15から成る。VPコントローラとVPポリシー・マネージャとの関係は、スレッド・コントローラとスレッド・ポリシー・マネージャとの関係と同種である、すなわちVPコントローラはVPポリシー・マネージャの依頼者である。VPコントローラがポリシーの決定を行うことが必要となった場合には必ず、VPコントローラはその決定を行うためにVP

ポリシー・マネージャをコールする。

【0183】物理的プロセッサはすべて同一のVPコントローラをランさせるが、それらは異なるVPポリシー・マネージャをランさせることができる。その結果、マルチプロセッサ・システムはシステムによる各物理的プロセッサの利用をカスタマイズすることが可能となる。また、システムは各物理的プロセッサ上で同じVPポリシー・マネージャをランさせることも可能である。

【0184】仮想マシンが抽象物理的プロセッサ上の仮想プロセッサをスケジュールしようとする場合、仮想マシンはその物理的プロセッサ上の仮想プロセッサ・コントローラをコールする。同様に、仮想マシンが、抽象物理的プロセッサから仮想プロセッサを除去しようとする場合には、仮想マシンはその物理的プロセッサ上の仮想プロセッサ・コントローラをコールする。各VPコントローラは、仮想プロセッサの状態変化を含め、その物理的プロセッサにマッピングされた仮想プロセッサを管理する。

【0185】VPポリシー・マネージャは、物理的プロセッサ上の仮想プロセッサのスケジューリングおよび移行に係わるすべてのポリシーの決定を行う。この決定には3つのタイプがある。第1に、VPポリシー・マネージャはVPからPPへのマッピングを決める。このマッピングは2つの異なるタイミングで行われる。すなわち、VPが最初にランされたときと、ブロックされていたVPが再びランされたときである。第2に、ポリシー・マネージャは、PP上のVPをランさせる順番と期間を決定する。最後に、VPポリシー・マネージャは、いつVPをあるプロセッサから他のプロセッサに移動(移行)させるべきかを決める。

【0186】これらの3つの決定によって、VPポリシー・マネージャはマシン上のワーク・ロードのバランスをとることができ、そして仮想マシンに関する物理的マシンの公正さに係わる性質を決めることができる。また、物理的プロセッサが故障したとき、故障許容VMのVPをどこに移動させるかを決定することができる。

【0187】スレッド・ポリシー・マネージャのようにVPはVPコントローラに対して良く定義されたインターフェースを提供する。VPポリシー・マネージャがそ

の決定を行うために用いるデータ構造はVPポリシー・マネージャに対して完全にプライベートである。これらのデータ構造は特定のVPポリシー・マネージャに対してローカルとできるか、またはVPポリシー・マネージャの種々の場合において共有できるか、またはそれらに対するアクセス手段を持たない。VPポリシー・マネージャは、スティングの異なる場合に対して異なる振舞をするようにカスタマイズすることができる。この機能により、スティングを、リアルタイム・システムや、会話型システムや、多量の計算を行うシステムなど、さまざまなオペレーティング・システム環境に対して、カスタマイズすることが可能となる。

【0188】最後に、スレッド・ポリシー・マネージャはスレッド間の負荷平衡および公正さに係わっているが、仮想プロセッサ・ポリシー・マネージャは、仮想マシン間および仮想プロセッサ間の負荷平衡および公正さに係わっている。

【0189】APM内の各物理的プロセッサは、仮想プロセッサ・コントローラ(VPC)と仮想プロセッサ・ポリシー・マネージャ(VPPM)を含んでいる。この点で、物理的プロセッサは構造的に仮想プロセッサと同一である。VPCは仮想プロセッサ上の状態変化に影響を与える。スレッドのように、仮想プロセッサはラン、レディー、ブロック、あるいは終了のいずれかの状態をとり得る。ラン状態のVPは物理的プロセッサ上で現在実行されている。レディー状態のVPはランすることが可能であるが、現在はランしていない。ブロック状態のVPは、なんらかの外部イベント(例えばI/O)を待っているスレッドを実行している。VPPMは物理的プロセッサ上のVPのスケジューリングを行う。そのスケジューリング・ポリシーはTPMが用いるものと同様である。VPPMは良く定義されたインターフェースをVPコントローラに対して提供する。異なるスティングのシステムは異なるVPポリシー・マネージャを備えることができる。

【0190】<例外の扱い>同期した例外および割込はスティングでは一様に扱われる。すべての例外には、例外を扱うための1組の動作を実行するハンドラーが関連している。ハンドラーはスレッド内で実行するプロシージャである。プロセッサP上で生じた例外は、Pの現在のスレッドのコンテキストを用いて実行する。スティングのマイクロ・カーネル内には特別な例外スタックは無い。プロセッサP上である例外(例えば、無効命令、メモリ保護破壊など)が生じた場合、Pの現在のコンティニュエーション(すなわち、プログラム・カウンタ、ヒープ・フロンティア、スタックなど)がまず退避される。次に例外ディスパッチャーは例外のターゲットを見つけるため、スレッドがランしている場合にはそれを中断し、そしてハンドラーのコンティニュエーションおよ

び引数をターゲット・スレッドのスタック上にプッシュする。次に、ディスパッチャーは(a)現在のスレッドを、単純にそれに復帰させることによって再開させるか、(b)ターゲット・スレッドを再開させるか、あるいは(c)このプロセッサ上の他のいずれかのスレッドを再開させるためにスレッド・コントローラをコールするか、いずれかを選択する。ターゲット・スレッドが再開された場合には、そのスレッドはそのスタックの最上のコンティニューエーションを実行する。これは例外ハンドラーのコンティニューエーションである。

【0191】スティングにおけるこの例外処理手段はいくつかの点で優れている。

【0192】1. この例外処理手段はプロシージャであるため、単にそれをコールするだけで例外を扱える。

【0193】2. 例外は、実行コンテキストを受け取るスレッドの実行コンテキストにおいて扱われる。

【0194】3. 例外は現在のスレッドのコンテキストにおいてディスパッチされる。

【0195】4. 一度ディスパッチされた例外はターゲット・スレッドの現在のコンティニューエーションとなり、そしてスレッドが再開されたとき自動的に実行される。

【0196】5. 例外はターゲット・スレッドが再開されたときのみ扱われる。

*【0197】6. 例外を扱うコードはスキームによって書かれ、そしてそのコードはコンティニューエーションとプロシージャを操作して所望の効果を達成する。

【0198】ファーストクラスのプロシージャとスレッド、明白なコンティニューエーション、動的な記憶領域の割り当て、ならびに均一なアドレッシング・メカニズムはすべてスティングの設計の中心的な特徴であり、その結果、スティングはこの例外のモデルを与えることが可能となっている。

10 【0199】同期した例外のターゲット・スレッドは常に現在のスレッドである。非同期的例外、すなわち割り込みはわずかに異なる形で扱われる。割り込みはどのスレッド(現在実行中のスレッドではない)でも制御できるので、このような例外を扱うためには、ハンドラーは、例外を直接処理するか、すなわち現在ランしているスレッドを中断して例外を扱うか、あるいは新しいハンドラーを生成する必要がある。割込ハンドラーもスキームのプロシージャであるため、ハンドラーを実行するためにスレッドを確立するか、あるいは現在のスレッドを用いる場合、単に適当なスレッドの現在のコンティニューエーションを、ハンドラーをコールするように設定すればよい。スティングの例外ディスパッチャーのための疑似コードを以下に示す。

* 【0200】

```
1: (define (exception-dispatcher type.
args)
2:   (save-current-continuation)
3:   (let ((target handler (get-target
&handler type args)))
4:     (cond ((eq? target (current-thread
)))
5:           (apply handler args))
6:     (else
7:       (signal target handler a
rgs)
8:       (case ((exception-priorit
y type))
9:         ((continue) (return))
10:        ((immediate) (switch-to
-thread target))
11:        ((reschedule) (yield-pr
ocessor))))))
```

ライン2では、現在のコンティニューエーションが現在のスレッドのスタックに退避される。このコンティニューエーションは、エスケープできず、そして一度だけコールされるので、上記スタックに上記コンティニューエーションを退避できる。ライン3では、ディスパッチャーが、例外の対象となるスレッドと、例外のタイプに対するハンドラーとを見つける。ライン4では、例外のターゲットが現在のスレッドであるかどうかチェックされ、そ

うなら、例外コンティニューエーションはプッシュされない(ライン5)。ディスパッチャーはハンドラーをむしろ単純にその引数に適用する。ディスパッチャーはすでに例外ターゲット(すなわち現在のスレッド)のコンテキストで走っているため、このことが有効である。例外のターゲットが現在のスレッドでない場合には、ディスパッチャーは例外をターゲット・スレッドに送る(ライン7)。スレッドに信号を送ることはスレッドを中断

し、信号ハンドラーとその引数を含むコンティニューエーションをスレッドのスタックにプッシュすること、そして信号ハンドラーが実行されるようにするスレッドを再開させることと等価である。ターゲット・スレッドに信号を送った後、ハンドラーはプロセッサ上で次にどのスレッドを走らせるかを定める(ライン8)。走らせるのはそれ自身の場合もあり(ライン9)、あるいはターゲット・スレッド(ライン10)か、または最も優先順位の高いスレッドの場合もある(ライン11)。

【0201】スティングの例外ハンドリング機能と他のオペレーティング・システムにおけるものとは、もう1つ重要な点で異なっている。例外を扱うスレッドは、システム内のユーザ・レベルのスレッドと違わないので(例えば、それらは自身のスタックとヒープを持っている)、また、例外ハンドラーは通常のファーストクラスのプロシージャであるため、ハンドラーは記憶領域を自由に割り当てることができる。ハンドラーによって生成されたデータは、他のデータが復元されるのと同じ方法で、ガーベッジ・コレクタによってリクレーンされる。例外ハンドリングのメカニズムと、より高レベルのスティングの抽象体との間の均一性のため、デバイス・ドライバを実現したとき、高い表現性および高い効率が得られる。このことは、上記均一性が無い場合には、並列言語あるいは並列オペレーティング・システムにおいて実現しない。

【0202】ファーストクラスのプロシージャとスレッド、明白なコンティニューエーション、動的な記憶領域の割り当て、ならびに均一なアドレッシング・メカニズムがすべてスティングの設計の特徴であるため、スティングはこの例外のモデルを与えることができる。

【0203】<並行パラダイム>以上、ソフトウェア・アーキテクチャについて詳しく説明したが、以下においてはいくつかの広範な並行パラダイムについて説明し、本発明のソフトウェア・アーキテクチャによってそれを実現する。

【0204】結果としての並行プログラムでは、並行して実行する各プロセスは、複合データ構造(例えば、アレーあるいはリスト)の値に影響を与える。または各プロセスは複合プロセスのグラフのメンバーである。プロセスの通信はこの結果の構造体またはグラフによる。そのcontributingプログラムがまだ評価中である結果の要素にアクセスを試みる表現は、プログラムが完了するまでブロックする。フューチャーは、結果としての並行アルゴリズムを実施するのに非常に適したオペレーションの良い例である。MultilisであるいはMult-Tの表現によって生成されたオブジェクト(フューチャーE)は、計算Eのためのスレッドを生成する。そしてリターンされたオブジェクトはフューチャーとして知られている。結果としてvを生じてEが終了したとき、フューチャーが確定したと言う。フュー

チャーにタッチする表現は、Eがまだ計算されている場合にはブロックし、他方、フューチャーが確立した場合にはvを与える。

【0205】図11に示す素朴なソーティング・プログラムでは、フューチャーの各例は新しいスレッドの生成を伴う。この振舞は望ましいものではない。それは、プロセス・ツリーのレベル1で計算を行うフューチャーはレベルi+1などにおいてその子に対して明らかなデータ依存性を有しているといった理由による。このプログラムにおいてデータ依存性があった場合、プロセッサおよび記憶装置の利用度が低下する結果となる。これは、生成されたライトウエイトのプロセスの多くが、まだ未評価のフューチャーのものとして他の値をリクエストするときブロックする必要があるか、または、例えば、小さい素数を計算するプロセスの場合、それらを生成するために必要なコストに比べ、少量の計算を行うためである。

【0206】スレッドの動的な状態は大きいオブジェクト(例えば、スタックおよびヒープ)から成るので、プロセスのブロッキングが頻繁に生じる場合、あるいはプロセスにグラニュラリティが小さすぎる場合、キャッシュおよびページのローカリティについては妥協する。

【0207】タッチおよびフューチャーの意味規則は、他のフューチャーGにタッチするフューチャーFは、Gがまだ確定していない場合、Gでブロックしなければならないということを命令する。T_FおよびT_EをそれぞれFおよびGのスレッド表現とする。Gでのタッチ・オペレーションのランタイム・ダイナミックスは、TBが(a)遅延またはスケジュールされたとき、(b)評価しているとき、(c)または確定したときのいずれかの場合、T_Eに対するアクセスを伴う場合がある。最後のケースでは、これらのスレッド間で同期化は不要である。ケース(b)の場合、T_FはT_Eが完了するまでブロックする必要がある。ケース(a)の場合、スティングでは重要な最適化を行う。これについては以下に説明する。

【0208】TFは、TG内に閉じ込められた閉包(Eと呼ぶ)を、コンテキスト・スイッチをブロックし、強制するより、むしろそれ自身のスタックヒープとを用いて評価することができる。実際、スティングでは、Eを通常のプロシージャとして扱い、Gのタッチを単純なプロシージャ・コールとして扱う。この場合、T_FがT_Eを吸収すると言う。T_Fは、その他の場合には必然的にブロックするという点でこの最適化は正しい。T_Fの動的なコンテキストを用いてEを適用することによって、T_Fが動作するVPは、コンテキスト・スイッチを実行するというオーバーヘッドを負わない。また、T_FのTCBが代りに用いられるので、T_Eに対してTCBを割り当てる必要がない。

【0209】この最適化は、コールしているスレッドが

必ずしもブロックする必要がない場合に用いられたとき、目立って異なった結果を導くのみとなる場合がある。例えば、 T_i が T_j によるスペキュラティブ・コールの要素であったとする。さらに、 T_i は分岐するが、他のスペキュラティブ・スレッド(T_j と呼ぶ)は分岐しないとす。吸収が無い場合には、 T_i および T_j は共に別々のスレッド・コンテキストを生む。しかし、吸収がある場合には、 T_i は T_j を吸収することができ、そして、 T_i がループするので T_j もループしよう。スレッドが吸収できるか、またはできない場合、ユーザはスレッドの状態をパラメータ化して、TCに通知することができる。スティングはこのためのインターフェース・プロシージャを提供する。

【0210】吸収のため、スティングはコンテキスト・スイッチングのオーバーヘッドを低減させ、そしてプログラムにおいてプロセスが互いに強いデータ依存性を示すとき、そのプログラムに対するプロセスのグラニュラリティを増大させる。もちろん、オペレーションを最も効果的なものにするため、スケジュールされたスレッドが吸収された状態になり得るよう、スレッドのグラニュラリティは十分に大きいものでなければならない。プロセスのグラニュラリティが小さすぎる場合には、吸収しているスレッドがそれらの値を要求できる前に、プロセスは吸収され得る可能性のあるスレッドの評価を開始しよう。

【0211】負荷にもとづくインライニングおよびレイジーなタスク生成は、他の並列Lispシステムに応用された2つの他の同種の最適化である。負荷にもとづくインライニングでは、現在のシステムの負荷がある特定のスレッショールドを越えた場合、スレッドはインライン(すなわち、吸収)される。この最適化では、プログラムの介入は不要であるだけでなく、ある種の条件のもとでは、本来終了するはずのプログラムがデッドロックあるいは長時間の停止状態になる場合がある。これはインライニングの決定が撤回できないからである。従ってこの最適化では、タスクが、そのデータ依存性のためにある順序で評価される必要があるとき、それとは異なる特定の評価の順序をタスクに課す。スレッドの吸収は、吸収されない場合にはスレッドがブロックするときのみ、そしてデータの依存性が保証されているときのみ生じるので、この問題の影響を受けない。

【0212】レイジーなタスクの生成は、負荷にもとづくインライニングに係わる多くの問題を解決する。レイジーなタスクの生成では、常にすべてのスレッドの評価がインラインされるが、しかしプロセスがアイドル状態となったとき、このインライニング・オペレーションを撤回可能とする。スレッドは実際に必要とされない限り決して生成されない。この設計ではプログラムの介入を必要とせず、本来デッドロックしないプログラムのデッドロックを招かず、そして、実際に発生されるタスク

の数が低減される。

【0213】スレッドの吸収はレイジーなタスクと主に2つの点で異なっている。(1)スレッドの吸収は、アプリケーションによって決まるスケジューリング・プロトコルが存在しても働く。レイジーなタスク生成はグローバルLIFOスケジュールと、インラインされたスレッドを保持するための単一の待ち行列の存在とを仮定する。(2)レイジーなタスク生成は、1つのプロセッサに対して1つのグローバル・ヒープを用いる。レイジーなタスクの生成では、タスクがスティールされたとき、スレッド吸収の場合よりローカリティは低下する。第2に、レイジーなタスク生成の場合のガーベッジ・コレクションでは、システム内のすべてのスレッドを停止させる必要がある(コレクタそれ自身が並列であっても)。スレッドの吸収の場合にはこの制約はない。

【0214】他の例はマスタ・スレーブのパラダイムであり、これは並列プログラムを構成するためのポピュラーな技術である。この技術では、発生されたプロセスのコレクションは先験的に行われる。マスタ・プロセスはいくつかのワーカー・プロセスを発生し、それらの結果を結合する。プロセスの通信は典型的には共有並行データ構造あるいは共有並行変数を通じて行われる。マスタ・スレーブ・プログラムがしばしば、ストック・マルチプロセッサ・プラットフォーム上の結果の並列プログラムより効率的である。それは、ワーカーが、それらの結果を発行する場合を除いて、ほとんど互いに通信する必要がないからである。そしてプロセスのグラニュラリティを調整でき、より高い性能が得られる。

【0215】スキームにおけるファーストクラスのダブル空間を最適化して実現するためにスティングを用いた。ダブル空間は、同期化コンテンツ・アクセサブル・メモリの抽象体として機能するオブジェクトである。ダブル空間は、マスタ/スレーブにもとづく多数のアルゴリズムを具体化するための自然の選択である。

【0216】ダブルはオブジェクトであり、ダブル・オペレーションはバインディング表現であって、ステートメントではないので、ファーストクラスの指示可能なダブル空間の存在により、モジュール性および表現性がさらに向上する。望ましい実施例では、ダブル空間は、同期化したベクトル、待ち行列、ストリーム、セット、共有変数、信号、あるいはバッグとして特殊化できる。ダブル空間上で許可されたオペレーションは、それらの表示において不変である。さらに、アプリケーションは必要ならダブル空間の間の継承階級を指定できる。

【0217】プロセスは新しいダブルをダブル空間に読み込んだり、除去したり、預けることができる。読み込みオペレーションあるいは除去オペレーションにおけるダブル・引数は“テンプレート”と呼ばれ、“?”を前に付けた変数を含むことができる。このような変数は“フォーマル”と呼ばれ、マッチ・オペレーションの結

果としてバインディング値を獲得する。これらのフォーマルによって獲得されたバインディング値は、下位の表現の評価において用いられる。従って、次のように書くことができる。

```
【0218】 (get TS [?x]
          (put TS [(+x1)]))
```

これによって1つのタプルがTSから除去され、1だけインクリメントされ、そしてTSに再び預けられる。

【0219】この実施例ではまた、スレッド吸収も利用して、タプル空間上で同期するグラニュラリティの細かい並列プログラムの構築を可能とする。スレッドはタプル内で真正な要素として用いられる。次の表現を実行するプロセスPを考える。

```
【0220】 (rd TS [x1 x2] E)
```

ここで、 x_1 と x_2 は非フォーマルである。さらに、TS内のタプルがオペレーション ($\text{spawn TS [E}_1 \text{ E}_2 \text{]}$) の結果として預けられているとする。このオペレーションは E_1 と E_2 を計算する2つのスレッド (TE_1 および TE_2 と呼ぶ) をスケジュールする。 TE_1 と TE_2 が共に完了すると、結果としてのタプルは2つの確定したスレッドを含んでいる。マッチング・プロシージャは、タプル内でスレッドに遭遇したとき、 thread-value を適用する。このオペレーションはそのスレッドの値を回収する。

【0221】しかし、Pが実行されるとき TE_1 がまだスケジュールされている場合には、Pはそれを自由に吸収でき、その結果が x_1 に一致するときは確定する。一致するものが存在しない場合には、Pは、スケジュールされた状態にあるかもしれない TE_2 を残して、他のタプルのサーチへと進む。その後、他のプロセスがこの同じタプルを調べることは可能であり、正当な理由があるなら TE_2 を吸収する。同様に、 TE_1 の結果が x_1 と一致するなら、Pは次に TE_2 を自由に吸収できる。 TE_1 または TE_2 のいずれかがすでに評価を行っている場合には、Pは、1つ（または両方）のスレッドでブロックするか、またはTS内で、他に一致する可能性のあるタプルを調べるかを選択する。タプル空間の意味規則は、この点でこの実施例に対して制約を課さない。

【0222】スティングの、ファーストクラスのスレッドとスレッド吸収との組み合わせは、共有データ構造を用いて、疑似要求によって駆動されるグラニュラリティの細かい（結果）並列プログラムを書くことを可能とする。この意味で、スレッド・システムは、構造にもとづく同期化（例えば、タプル空間）とデータフロー・スタイルの同期化（例えば、フューチャー／タッチ）との間の意味のある区別の最小化を試みる。

【0223】スベキュラティブ並列は重要なプログラミング技術であるが、それを実現した際に生じるランタイムのオーバーヘッドのために、しばしば効果的に用いることができない。スベキュラティブ・プログラミング・

モデルをサポートするシステムに最も頻繁に係わる2つの特徴は、他のものより一層有望なタスクを奨励する能力と、不要な計算を中止および再利用（そして、恐らく取消し）する手段を有することである。

【0224】スティングは次のことによって、プログラマがスベキュラティブ・アプリケーションを書くことを可能とする。

【0225】1. ユーザがスレッドの優先順位を明示的にプログラムすることを可能とする。

【0226】2. 他のスレッドが完了したとき、あるスレッドがウェイトできるようにする。

【0227】3. スレッドが他のスレッドを終了させることを可能とする。

【0228】優先順位をプログラムできるので、有望なタスクはそうでないものより先に実行することができる。タスクの組の中で最初に終了するタスク α は、その終了の際、ブロックされているスレッドをどれでも目覚めさせることができる。この機能によって、スティングはOR並列の有用な形態をサポートできる。タスク α は、そのタスクの組の中の他のタスクはすべて、それらの結果が不要であると確定されたなら、終了させることができる。しかし、スティングを用いた理論的計算は、不要なタスクによってもたらされたノンローカルな副作用を取消すことはできないであろう。このシステムは基本的な逆戻りのメカニズムは提供しない。

【0229】 wait-for-one コンストラクトを実現することを考える。このオペレータは、並行してこの引数のリストを評価し、その最初の引数によって生成された値を復帰させ、終了する。従って、表現 ($\text{wait-for-one } a_1 a_2 \dots a_i \dots a_n$) において a_i から v が生じた場合、この表現は v を復帰させ、そして、プログラマが必要とするなら、残っているすべての a_j 、 $j \neq i$ の評価を終了する。

【0230】AND並列を実現した wait-for-all コンストラクトの仕様の同様である。これも並行してその引数を評価する。ただしすべての引数を終ったときのみ真を復帰させる。従って表現 ($\text{wait-for-all } a_1 a_2 \dots a_i \dots a_n$) は、この表現を実行するスレッドはすべての a_i が終るまでブロックされているので、障壁同期化ポイントとして機能する。このオペレーションの実現は、スベキュラティブ wait-for-one オペレーションの実現と非常に似ている。

【0231】TCはこれらのオペレーションを、共通プロシージャである block-on-set を用いて実現する。スレッドおよびTCBは、この機能をサポートするように定義されている。例えば、TCB構造体に関連しているのは、TCBの関連するスレッドが再開できる前に終了しなければならぬ、グループ内のスレッドの数に関する情報である。

【0232】block-on-setは、スレッドのリストとカウントを取る。これらのスレッドは、上述したwait-for-oneオペレーションおよびwait-for-allオペレーションの引数に対応している。カウントの引数は、現在のスレッド（すなわち、block-on-setを実行しているスレッド）が再開を認められる前に終了しなければならないスレッドの数を表している。この数が1の場合、結果はwait-for-oneを実現したものであり、上記数がnの場合、結果はwait-for-allの実現である。

【0233】組の中のスレッド T_i と、 T を待つべき現在のスレッド（ T_i ）との関係は、下記のものに対する参照を含むデータ構造（スレッド・バリア（TB）と呼ばれる）内で維持される。

【0234】1. T_i のTCB
2. T_i 上でブロックされている他のウエイターのTB（存在する場合）block-on-setを定義するプログラムを、図12に示す。

【0235】次のコール

(block-on-set m T_1 T_2 ... 20 T_n)

```

*
(define (wait-for-one . block-group)
  (block-on-group 1 block-group)
  (map thread-terminate block-group)
)
```

T がwait-for-oneを実行する場合、それはblock-group引数内のすべてのスレッド上でブロックする。 T が再開されるとき、 T は、利用できるいずれかの仮想プロセッサのTPM内のレディー待ち行列に配置される。 T の再開のとき実行されるマップ・プロセスは、そのグループ内のすべてのスレッドを終了させる。

【0239】スティングのプロシージャwait-for-allは、このオペレーションを省略できる。それは、そのブロック・グループ内のすべてのスレッドは、このオペレーションを実行するスレッドが再開される前に、終了することが保証されているからである。

【0240】スティングは、8プロセッサのSilicon Graphics Power Series (MIPS R3000)と、16プロセッサのSilicon Graphics Challenge (MIPS R4400)の両方において実現した。両マシンは、共有（キャッシュ・コヒーレント）マルチプロセッサである。この抽象物理的マシン構成では、物理的プロセッサはライトウエイトのUnixスレッドにマッピングされる。マシン内の各プロセッサは、このようなスレッドの1つをランさせる。

【0241】以上、コンピュータ・ソフトウェア・アーキテクチャの望ましい実施例について記述し、説明したが、当業者にとって明らかなように、本発明の広範な原

*は現在のスレッド（ T と言う）に、 m 個の T_i （ $m \leq n$ ）が終了したときアンブロックさせる。これら T_i のそれぞれは、それらのウエイターのチェーン内に T に対する参照を有している。

【0236】アプリケーションはblock-on-setを、アプリケーションが終了したとき a_i によって起動されるプロシージャwake-up-waitersと共に用いる。wake-up-waitersは、そのスレッド引数内のウエイター・スロットから、連鎖状のウエイターのリストを調べる。ウエイター数がゼロになるウエイターは、いずれかのVPのレディー待ち行列に挿入される。TCは、スレッド T が終了したときはいつもwake-up-waitersを起動する（例えば、 T が終了したとき、または異常に存在するときはいつも）。 T の終了を待っているスレッドは、すべてこのようにしてリスケジュールされる。

【0237】これは2つのプロシージャが与えられると、wait-for-oneは次のように簡単に定義することができる。

【0238】

理および趣旨から逸脱することなく、種々の変形や変更を加えることは可能である。

【0242】

【発明の効果】以上説明したように本発明によれば、高度並列マルチプロセッサ／マルチコンピュータ・システムを制御するための、現代のプログラミング言語に対する非常に効率の良いサブストレートとして役立つコンピュータのオペレーティング・システム・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0243】更に本発明によれば、カスタマイズ可能な仮想マシンにもとづく非同期の計算のためのソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0244】また本発明によれば、仮想プロセッサ上でファーストクラスのオブジェクトとしてライトウエイト・スレッドをサポートするソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0245】更に本発明によれば、カスタマイズ可能なポリシー・マネージャを、特にユーザ・レベルを含むソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0246】また、本発明によれば、カスタマイズ可能な仮想トポロジーを含むソフトウェア・アーキテクチャ

を用いた高度並列コンピュータ・システムの制御方式が得られる。

【0247】更に本発明によれば、スレッド吸収、遅延TCB割り当て、ならびに記憶装置共有の場所としてのスレッド・グループを含むソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0248】また本発明によれば、多様な形態のポートを含むソフトウェア・アーキテクチャを用いた高度並列コンピュータ・システムの制御方式が得られる。

【0249】更に本発明によれば、上述のようなソフトウェア・アーキテクチャを用いて制御されるコンピュータ・システムが得られる。

【図面の簡単な説明】

【図1】本発明の一実施例によるソフトウェア・アーキテクチャを用いた制御方式を示すブロック図である。

【図2】図1の抽象物理的マシンおよび仮想マシンを示す図である。

【図3】本発明のオペレーティング・システムの抽象アーキテクチャを示す概略ブロック図である。

【図4】本発明で用いるスレッドの状態およびTCBの状態の遷移を示す図である。

【図5】本発明で用いる記憶装置の構成を表す概略図である。

【図6】本発明で用いるスレッドのプログラミングを説明するためのプログラムを示す図である。

【図7】本発明で用いる物理的プロセッサの2Dメッシュ上で多重化された仮想プロセッサの3Dメッシュを生成するプログラムを説明するための図である。

【図8】本発明で用いるコンテキスト・スイッチを開始するプログラムを示す図である。

【図9】本発明で用いるコンテキスト・スイッチを終了するプログラムを示す図である。

【図10】本発明で用いる新しいスイッチを開始するプログラムを示す図である。

【図11】本発明で用いるグラニュラリティの細かい適

応並列ソート・アルゴリズムのための最上位のプロシージャのプログラムを示す図である。

【図12】本発明で用いるblock-on-setを定義するプログラムを示す図である。

【符号の説明】

10	抽象物理的マシン
11	物理的トポロジー
12	抽象物理的プロセッサ
13	仮想プロセッサ・コントローラ
14	仮想マシン
15	仮想プロセッサ・ポリシー・マネージャ
16	仮想プロセッサ
17	スレッド・コントローラ
18	スレッド
19	スレッド・ポリシー・マネージャ
20, 20'	仮想トポロジー
24	仮想マシン/アドレス空間
26	グローバル記憶プール
28	グローバル共有オブジェクト
30	ルート環境
31	スタック
32	TCB
33	ローカル・ヒープ
35	グローバル・ヒープ
36	遅延
38	スケジューラ
40	評価
42	吸収
44	確定
46	初期化
48	レディー
50	ラン
52	ブロック
54	保留
56	終了

【図6】

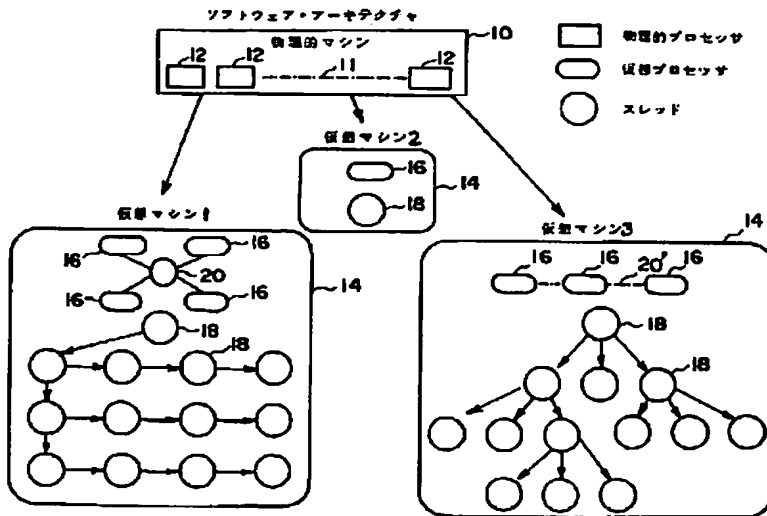
```
(define (filter op n input)
  (let loop ((x (hd input))
             (output (make-stream))
             (last? true))
    (cond ((zero? (mod x n))
           (loop (rest input) output last?))
          (else?
           (op (lambda () (filter op x output)))
             (loop (rest input) (attach x output) false)))
    (else (loop (rest input) (attach x output) last?))))

(define (sieve op n)
  (let ((input (make-integer-stream n)))
    (op (lambda () (filter op 2 input)))))
```

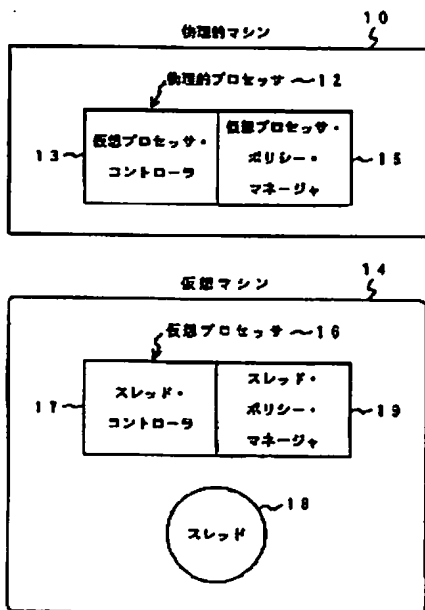
【図7】

```
(define (create-3D-mesh depth)
  (let ((pw (get-pw-width))
        (ph (get-ph-height))
        (3D-mesh (make-array w h depth)))
    (let -i- ((i 0))
      (if (< i pw-width)
          (let -j- ((j 0))
            (if (< j ph-height)
                (let ((vp (create-vp (get-pw i))))
                  (set-vp-address vp (vector i j))
                  (let -k- ((k 0))
                    (cond ((< k depth)
                           (set-aref! 3D-mesh vp)
                           (-+ + k 1))))
                        (-+ + j 1))))))
          (-+ + i 1))))))
```

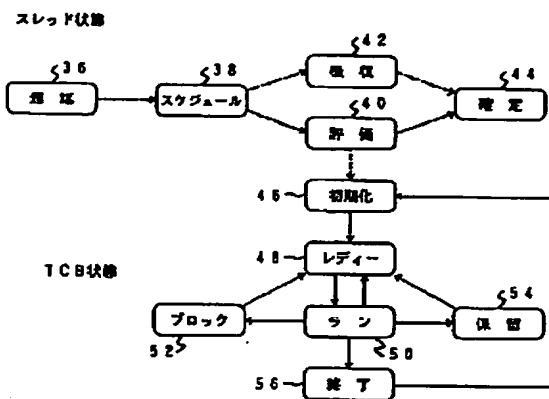
【図1】



【図2】



【図4】



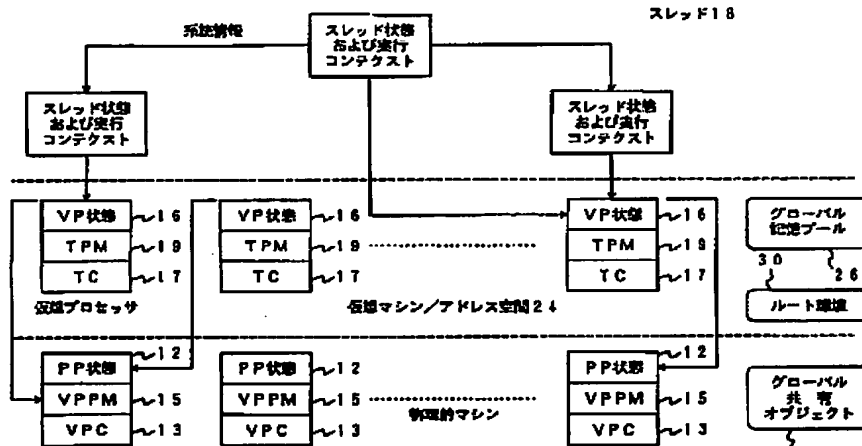
【図9】

```

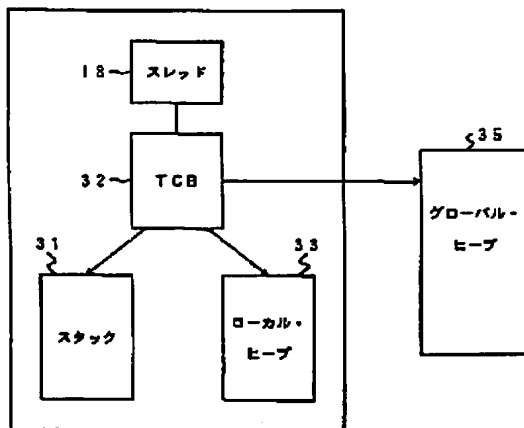
(define (finish-context-switch)
  (let* ((vp (current-vp))
        (previous (vp-current-tcb vp))
        (state (tcb-state previous)))
    (set-vp current-tcb vp (current-tcb))
    (cond ((not? previous (current-tcb))
          (cond ((tcb-state-ready? state)
                 (if (not? previous (vp-root-tcb vp))
                     (top-queue-ready-thread vp previous)))
                ((tcb-state-blocked? state)
                 (thread-state-release (tcb.thread previous)))
                ((tcb-state-suspended? state)
                 (top-queue-suspended-thread vp (tcb.thread previous)))
                (thread-state-release (tcb.thread previous))))))
    (enable-promotion)))

```

【図3】



【図5】

記憶装置の構成
スレッド動的コンテキスト

評価中のスレッドオブジェクト

【図10】

```

(define (start-new-thread)
  (let ((x (error-value "Thread has no value")))
    (unwind-protect (set z (catch exit
      (set-tcb-uid-handler (current-tcb) exit)
      ((thread-thunk (current-thread))))))
      (let ((thread (current-thread)))
        (thread-go thread)
        (set-thread-value thread z)
        (makeup-waiters thread)
        (set-tcb-state (current-tcb) tcb-state/dead)
        (start-context-switch tcb-state/dead))))))

```

【図8】

```

(define (start-context-switch next-state)
  (disable-preemption)
  (let* ((vp (current-vp))
        (next (let ((tcb (tcb-get-next-thread vp)))
          (if (false? tcb)
              ((tcb-state-ready? next-state)
               (current-tcb)
               (tcb-vp-idle vp))
              tcb))))))
    (cond ((eq? next (current-tcb)))
          ((tcb? next)
           (set-tcb-state next (tcb-state/running)
             (set-thread-vp (tcb-thread next) vp)
             (cond ((eq? next-state tcb-state/dead)
                    (return-current-tcb-to-vp-pool vp)
                    (restore-tcb-end-registers next))
                 (else
                  (save-current-tcb-registers)
                  (restore-tcb-end-registers next))))))
          ((thread? next)
           (thread-mutex-acquire next)
           (let ((tcb (allocate-tcb next-state vp)))
             (setup-new-thread next vp tcb)
             (if (next? tcb (current-tcb))
                 (save-current-tcb-registers)
                 (thread-mutex-release next)
                 (start-new-tcb tcb start-new-thread))))))

```

【図11】

```

(define (pbisort root spare up?)
  (cond ((node-left root)
        (let ((left-half (future (pbisort (node-left root) root up?))))
          (pbisort (node-right root) spare (not up?))
          (touch left-half)
          (pbimerge root spare up?))
        (else (compare-and-swap root spare up?))))

(define (pbimerge root spare up?)
  (let loop ((root root) (spare spare))
    (if (compare-and-swap root spare up?)
        (flip-tree-1 root up?)
        (flip-tree-2 root up?))
    (cond ((node-left root)
          (let ((left-half (future (loop (node-left root) root))))
            (loop (node-right root) spare)
            (touch left-half))))

```

【図12】

```

(defina (block-on-wait count threads)
  (thread-mutex-acquire (current-thread))
  (let loop ((threads threads)
              (cnt count))
    (cond ((fx-zero? cnt)
           (thread-mutex-release (current-thread)))
          ((null? threads)
           (set-tcb-wait-count (current-tcb) cnt)
           (tcb-state->blocked (current-thread)))
          (else (let ((thread (car threads)))
                   (thread-mutex-acquire thread)
                   (cond ((thread-determined? thread)
                          (thread-mutex-release thread)
                          (loop (cdr threads) (fx-cnt 1)))
                         (else (let ((tb (make-tb)))
                                 (set-tb-waiter tb (current-thread))
                                 (set-tb-thread tb thread)
                                 (set-tb-next tb (thread-waiters thread))
                                 (set-thread-waiters thread tb)
                                 (thread-mutex-release thread)
                                 (loop (cdr threads) cnt))))))))))

```